

应用统计与 SPSS 应用

朱红兵 编著

卢纹岱 审校

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

本书以综合性的实例为前提,根据研究问题的类型,着重讲述统计方法的选择和统计结果的解释等实际应用。全书包括 12 章,内容包括 SPSS 的基本功能与统计方法的选择,数据资料的整理与描述,常见的几种概率分布,参数估计,常用的概率抽样方法,总体参数的假设检验,非参数假设检验,多因素方差分析,相关与回归分析,聚类与判别分析,因子分析与主成分分析等内容。为便于读者学习,本书所附光盘中包含书中所有例题数据。

本书例题丰富,实用性强,提供正确使用统计的方法及对统计结果的科学解释,是非统计专业本科生和研究生首选的实操教材,也是统计学专业的学生更好地使用软件解决实际问题必不可少的参考书。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

图书在版编目(CIP)数据

应用统计与 SPSS 应用 / 朱红兵编著. —北京:电子工业出版社, 2011.1

ISBN 978-7-121-12760-1

I. ①应… II. ①朱… III. ①统计分析—软件包, SPSS—高等学校—教材 IV. ①C819

中国版本图书馆 CIP 数据核字(2011)第 004302 号

策划编辑:杨丽娟

责任编辑:杨丽娟

印 刷:北京市顺义兴华印刷厂

装 订:三河市双峰印刷装订有限公司

出版发行:电子工业出版社

北京市海淀区万寿路 173 信箱 邮编:100036

开 本:787×980 1/16 印张:44.5 字数:974 千字

印 次:2011 年 1 月第 1 次印刷

印 数:4000 册 定价:69.00 元(含光盘)

凡所购买电子工业出版社图书有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系。
联系及邮购电话:(010) 88254888。

质量投诉请发邮件至 zlt@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线:(010) 88258888。

前 言

我们所处的时代是信息时代，信息时代离不开在大量的信息中去获取科学有用的信息，由于采集信息的方法中，很多时候使用了统计中随机采样的方式，因而使得收集到的信息中无不戴上随机的烙印，这使得对采集到的信息进行处理的数学方法也离不开统计学的身影。毫无疑问，统计学的理论和方法为处理自然科学和社会科学研究中众多受随机因素影响的实际问题，提供了有力的工具。

需求决定供给，正是这种时代的要求，目前，应用统计已成为各大学为许多非统计专业的本科生和研究生开设的一门必修课程。

众所周知，统计理论和方法在应用过程中的瓶颈之一，是其繁杂的计算过程，为了提高计算效率，统计学家不得不在简化计算方面花费很多精力，也由此产生了许多有助计算简化的算法，这些内容在传统的应用统计书籍中都占据了大量的篇幅，也需要占用读者很多宝贵的时间来掌握。而今随着计算机的普及专用统计软件的使用，这些影响统计普及和应用的障碍已不复存在。

本书是针对非统计专业的本科生、研究生及需要用统计方法来处理社会各领域科研问题的读者而编写的。本书以当今国际上最流行的统计软件之一 **SPSS** 为平台，以介绍统计概念、方法在实际中的应用和正确使用 **SPSS** 中的众多统计程序为立足点，目的是通过对初级、中级统计学知识的介绍，帮助非统计专业的读者来学习如何正确使用统计方法、如何判定这些方法与 **SPSS** 中程序的对应关系，以及如何正确分析和解释 **SPSS** 的输出结果。

因此，对各种统计方法的证明过程不作为本书的重点，多数都轻描淡写一带而过。全书侧重于应用，突出实用性，书中列举的大量例题均来源于实际科研中，通过对这些实例的解析，来帮助读者达到对书中所给出的各种统计方法的理解。

因在卢纹岱主编的《**SPSS** 统计分析（第 1-4 版）》书中，已对 **SPSS** 软件的基本操作方法及各种选项的解释上做了详细的阐述，因此，本书对这些方面不再作详细介绍，而把 **SPSS** 中的操作重点放在究竟选何种选项上，即如何正确选择适宜的统计方法上。

本书可作为非统计专业的本科生和研究生的应用统计教材和教学参考书，也可作为从事数据分析或统计应用的各领域、各专业研究人员的统计工具书。

本书共分 12 章。前 7 章为常用统计部分，主要介绍统计方法的选择、抽样方法、数据资料的收集和整理、常用的分布和统计推断的方法，它适合于非统计专业的本科生作

为统计入门课程内容。后 5 章偏向多元统计分析，主要介绍多因素方差分析、正交试验设计及其分析、相关与回归分析、聚类与判别分析、主成分、因子分析与对应分析，它适合于非统计专业的研究生作为统计入门课程内容。

在本书的编写过程中，卢纹岱教授不但担任了本书的审校，而且自始至终为本书的成稿提出了许多有益的建议和热情的鼓励和帮助，在此深表谢意。

全书由朱红兵编著。在编写过程中，苏林、朱启钊、苏为夏、林建亭、苏玉成、宋阳等同志在资料收集整理、数据录入、绘图、核对等方面做了大量工作，在此深表谢意。

由于编者的水平有限，错误之处在所难免，敬请读者批评指正。

反馈意见，请发电子邮件至：zhuhongbing@cipe.net.cn。

最后，尤其要感谢我的爱人苏林和我的家人，让我在没有家庭负担的情况下顺利完成本书的撰写工作。

编 者

2011 年 1 月于北京

目 录

第 1 章	SPSS 的基本功能与统计方法的选择	1
1.1	SPSS 主要功能概述	2
1.1.1	SPSS 的主菜单简介	2
1.1.2	SPSS 的主要统计分析功能	3
1.1.3	SPSS 的菜单与应用统计分析工作的主要步骤的对应关系	4
1.2	研究设计	6
1.2.1	研究指标的选择与设计	7
1.2.2	研究对象的确定	7
1.2.3	抽样设计	8
1.2.4	样本容量的确定	9
1.2.5	实验设计中用到的一些基本术语	10
1.2.6	常用的实验设计	11
1.2.7	在研究设计中的统计分析预案	13
1.2.8	在研究设计中需要用到 SPSS 中的子菜单、过程和程序	15
1.3	整理数据资料	15
1.3.1	在 SPSS 中建立数据文件	15
1.3.2	在 SPSS 中读取数据文件	16
1.3.3	在 SPSS 中合并数据文件	16
1.4	分析数据资料	17
1.4.1	研究目的是对数据资料作一般性描述	17
1.4.2	研究目的是对不同总体在相同指标上是否有差异进行推断	23
1.4.3	研究目的是检查两个或多个变量之间的关联或相关	25
1.4.4	研究目的是缩减指标	26
第 2 章	数据资料的整理与描述	28
2.1	SPSS 数据文件的制作	28
2.1.1	统计资料的类型与变量类型、测度类型的对应关系	29
2.1.2	建立调查问卷的数据文件	32

2.1.3	将 EXCEL 中建立的数据文件变成 SPSS 中数据文件	40
2.1.4	数据文件的合并	43
2.1.5	数据文件的转置和重新构建	48
2.1.6	其他特殊数据文件的建立	53
2.2	数据资料的整理：频数分布表的制作	59
2.2.1	定性数据资料频数分布表的制作	59
2.2.2	定量数据资料频数分布表的制作	60
2.3	数据资料的描述	65
2.3.1	统计图	65
2.3.2	统计表	80
2.3.3	统计量	84
2.3.4	分布形态	97
2.4	探索分析	99
2.4.1	探究分析的意义	99
2.4.2	实例分析	99
2.5	计算派生指标	110
2.6	排名	111
2.7	分析多重应答集	113
2.7.1	多选题的处理	113
2.7.2	排序题的处理	117
第 3 章	常见的几种概率分布	121
3.1	事件和概率	121
3.1.1	事件	121
3.1.2	事件之间的关系和运算	122
3.1.3	事件的频率和概率	123
3.2	随机变量和概率分布	126
3.2.1	随机变量	126
3.2.2	离散型随机变量的概率分布	127
3.2.3	连续型随机变量的分布	135
第 4 章	参数估计	153
4.1	参数的点估计	153
4.1.1	参数的矩估计法	154
4.1.2	参数的极大似然估计法	159
4.1.3	估计量的评选标准	162

4.2	参数的区间估计	167
4.2.1	区间估计的概念	167
4.2.2	正态总体均值的置信区间	168
4.2.3	正态总体方差的置信区间	171
4.2.4	两个正态总体均值差和方差比的区间估计	173
4.2.5	非正态总体参数的近似区间估计	182
4.2.6	其他总体参数及参数的区间估计	186
4.2.7	估计值的误差限及估计精度	190
第 5 章	几种常用的概率抽样方法	193
5.1	抽样概述	193
5.2	简单随机抽样	194
5.2.1	样本容量的确定	195
5.2.2	简单随机抽样过程	199
5.2.3	简单随机抽样的估计	213
5.3	系统随机抽样	224
5.3.1	系统随机抽样概述	224
5.3.2	系统随机抽样在 SPSS 中的实现	224
5.3.3	系统随机抽样的估计	225
5.4	PPS 抽样	227
5.4.1	PPS 抽样概述	227
5.4.2	PPS 抽样在 SPSS 中的实现	227
5.4.3	PPS 抽样的估计	231
5.5	PPS Brewer 抽样	232
5.5.1	PPS Brewer 抽样概述	232
5.5.2	PPS Brewer 抽样在 SPSS 中实现	233
5.5.3	PPS Brewer 抽样的估计	236
5.6	分层随机抽样	237
5.6.1	样本容量的确定	237
5.6.2	分层随机抽样过程	239
5.6.3	分层随机抽样的估计	242
5.7	整群抽样	244
5.7.1	整群抽样概述	244
5.7.2	整群抽样在 SPSS 中的实现	245
5.7.3	整群抽样的估计	245

5.7.4 整群抽样的实例分析	246
5.8 多阶抽样	248
5.8.1 多阶抽样概述	248
5.8.2 多阶抽样实例分析	249
第 6 章 假设检验	255
6.1 假设检验概述	255
6.1.1 何谓统计假设	255
6.1.2 可否直接根据试验结果数据值大小来做出拒绝或不拒绝统计假设的结论	255
6.1.3 何谓统计检验	259
6.1.4 假设检验的种类	260
6.1.5 假设检验中易犯的两类错误	261
6.2 一元正态总体均值差异的显著性检验	262
6.2.1 单样本 t 检验	262
6.2.2 独立样本 t 检验	264
6.2.3 配对样本 t 检验	270
6.2.4 单因素方差分析	274
6.3 多元正态总体均值差异的显著性检验	292
6.3.1 多元正态分布基本概述	292
6.3.2 多元正态总体均值差异的检验方法	294
6.3.3 多个协方差阵相等检验—Box's M 检验	296
6.3.4 随机误差的独立性检验—Bartlett 球型检验	297
6.3.5 实例分析	298
6.4 非正态总体参数的假设检验	303
6.4.1 非正态总体的均值检验	303
6.4.2 指数分布总体参数的检验	306
第 7 章 非参数假设检验	308
7.1 二项分布检验	308
7.1.1 二项分布检验概述	308
7.1.2 二项分布检验实例分析	308
7.2 卡方拟合分布检验	310
7.2.1 对多项分布各项概率已知时卡方拟合分布检验	310
7.2.2 对多项分布各项概率未知时的卡方拟合分布检验	313
7.3 序列随机性的游程检验	317
7.3.1 游程检验概述	317

7.4 柯尔莫哥洛夫-斯米诺夫检验	322
7.4.1 柯尔莫哥洛夫-斯米诺夫检验基本概述	322
7.4.2 实例分析	323
7.5 两个独立样本的检验	324
7.5.1 曼-惠特尼 U 检验和威尔科克森秩和检验	325
7.5.2 柯尔莫哥洛夫-斯米诺夫检验 Z 检验	328
7.5.3 摩西极端值反应检验	330
7.5.4 沃尔德-乌尔夫威兹游程检验	332
7.6 多个独立样本的检验	333
7.6.1 克鲁斯卡-沃里斯 H 检验	334
7.6.2 中位数检验	336
7.6.3 乔卡契尔-特普斯特拉检验	338
7.7 两个相关样本检验	340
7.7.1 威尔科克森检验	340
7.7.2 符号检验	342
7.7.3 麦内玛检验	344
7.7.4 边缘同质检验	345
7.8 多个相关样本检验	347
7.8.1 弗里德曼检验	347
7.8.2 肯德尔调和系数 (Kendall's W) 检验	349
7.8.3 克科伦 Q 检验	351
7.9 交叉表资料的检验	352
7.9.1 二维交叉表资料的独立性检验	354
7.9.2 多维交叉表资料的条件独立性和齐性检验	393
第 8 章 多因素方差分析和协方差分析	402
8.1 单因变量单因素嵌套设计中的方差分析	403
8.1.1 单因变量单因素嵌套设计的基本概述	403
8.1.2 单因变量单因素嵌套设计实例分析	405
8.2 单因变量单因素随机区组设计中的方差分析	410
8.2.1 单因变量单因素随机区组设计的基本概述	410
8.2.2 单因变量单因素随机区组设计实例分析	412
8.3 单因变量多因素试验的方差分析	416
8.3.1 单因变量双因素完全随机试验的方差分析	416
8.3.2 单因变量三因素完全随机试验的方差分析	449

8.4	单因变量协方差分析	468
8.4.1	单因变量协方差分析基本概述	468
8.4.2	单因变量协方差分析的实例分析	470
8.5	重复测量资料的方差分析	472
8.5.1	重复测量资料方差分析的基本概述	472
8.5.2	重复测量资料实例分析	476
第 9 章	正交试验设计与数据分析方法	481
9.1	正交试验设计方法的优点和特点	482
9.1.1	正交表	482
9.1.2	正交试验设计方法	485
9.2	正交试验设计的基本步骤	488
9.3	正交试验设计实例	489
9.4	正交试验设计的极差分析	497
9.4.1	极差分析的基本步骤	497
9.4.2	极差分析法的实例分析	497
9.5	正交试验设计的方差分析	511
9.5.1	正交试验设计方差分析的基本原理	511
9.5.2	正交试验设计方差分析实例	512
第 10 章	相关与回归分析	515
10.1	线性与趋势性相关分析	515
10.1.1	Pearson (皮尔逊) 相关系数	516
10.1.2	Spearman (斯皮尔曼) 秩相关	520
10.1.3	Kendall's tau-b (肯德尔 τ -b) 相关系数	523
10.2	偏相关分析	526
10.2.1	偏相关的概念	526
10.2.2	偏相关实例分析	527
10.3	距离分析	528
10.3.1	距离分析概述	528
10.3.2	距离分析的实例分析	536
10.4	典型相关	539
10.4.1	典型相关分析的数学模型	540
10.4.2	典型相关系数的检验	541
10.4.3	冗余测度	542
10.4.4	实例分析	543

10.5	线性回归分析	551
10.5.1	线性回归分析概述	551
10.5.2	一元线性回归分析	552
10.5.3	曲线估计-一元非线性回归	577
10.5.4	多元线性回归分析	588
10.6	逻辑斯蒂回归分析	607
10.6.1	逻辑斯蒂回归分析概述	607
10.6.2	二元逻辑斯蒂回归分析	607
10.6.3	多项逻辑斯蒂回归分析	615
第 11 章	聚类分析与判别分析	625
11.1	聚类分析	625
11.1.1	聚类分析的作用	625
11.1.2	聚类分析中常用的统计量	626
11.1.3	系统聚类法	626
11.1.4	典型指标的选择	637
11.1.5	动态聚类分析	638
11.1.6	判别分析	642
第 12 章	主成分分析、因子分析与对应分析	665
12.1	主成分分析	665
12.1.1	主成分分析及其基本思想	665
12.1.2	主成分分析的数学模型及求法	666
12.1.3	主成分的性质	668
12.1.4	主成分的应用及其注意点	668
12.1.5	主成分实例分析	669
12.2	因子分析	673
12.2.1	因子分析的数学模型及模型系数的统计意义	673
12.2.2	因子载荷矩阵的估计	675
12.2.3	因子旋转	676
12.2.4	因子得分	679
12.2.5	实例分析	680
12.3	对应分析	685
12.3.1	对应分析的基本原理	685
12.3.2	对应分析实例分析	689
参考文献		700

第 1 章 SPSS 的基本功能与统计方法的选择

SPSS 原意为 Statistical Package for the Social Sciences，即“社会科学统计软件包”，它是一个组合式软件包，集数据整理、科学计算、分析过程和结果输出等功能于一身。于 20 世纪 60 年代末由美国斯坦福大学的三位研究生研制。1984 年 SPSS 首先推出了世界上第一个统计分析软件微机版本 SPSS/PC+，是世界上最早的统计分析软件，在国际学术界有一条不成文的规定，即在国际学术交流中，凡是用 SPSS 软件完成的计算和统计分析，可以不必说明算法，享有极高的声誉。随着公司的进一步发展，SPSS 公司已于 2000 年正式将英文全称更改为 Statistical Product and Service Solutions，意为“统计产品与服务解决方案”。它是一款在调查统计行业、市场研究行业、医学统计、政府和企业的数据分析应用中久享盛名的统计分析工具。已广泛应用于自然科学、技术科学、社会科学的各个领域。随着统计学的不断发展，SPSS 的功能也得到了进一步的拓展，已有迹象表明，SPSS 将更改其名，用“PASW”取代“SPSS”。PASW 英文全称为“Predictive Analytics Software”，即预测分析软件。为使读者不至于混淆本书中所涉及的内容，因此，值得一提的是，本书是基于 SPSS16.0 基础上撰写的。

本章的主要内容是根据科研工作中常用的统计分析的一般工作步骤，建立起与 SPSS 之间的桥梁，即对 SPSS 的基本功能与常用的 80% 的统计方法进行必要的归纳总结，欲使之成为在实际统计分析工作中选择 SPSS 的基本功能与统计方法的向导，所以，必然需要用到后面章节中的许多知识。因此，对于尚未掌握统计基础的读者，更不必惊慌，不要急于弄清楚本章中所有内容，只需在掌握了本章中的一些基本知识和概念后，就完全可以跳过对方法的选择归纳中的大部分内容，等在对后面几章的内容有了大致的理解，以及建立起一些基本的统计知识和 SPSS 的基本操作过程后，再回过头来关注本章其余内容也不迟。但对于已有一定科研经验和统计基础的读者，通过阅读本章归纳性的总结，或许能加快找到解决科研中遇到的问题所对应的统计方法，以达到事半功倍的效果，这正是作者所期望的。

1.1 SPSS 主要功能概述

1.1.1 SPSS 的主菜单简介

在 SPSS 中，菜单栏共有 11 个选项，见图 1-1。分别是：

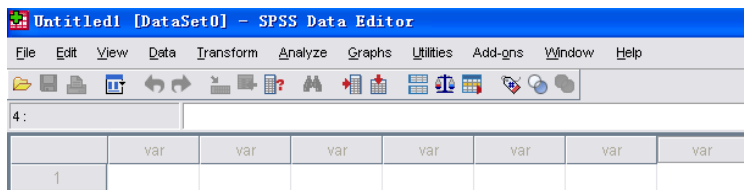


图 1-1 SPSS16.0 中的主菜单

1. **File:** 文件管理菜单，有关文件的建立、调入、存储、显示和打印等。
2. **Edit:** 编辑菜单，有关文本内容的选择、复制、剪贴、寻找和替换等。
3. **View:** 视窗菜单，有关 SPSS 数据编辑窗口外观、工具条显示、数据视窗和变量视窗的转换、单元格线显示、关闭等。
4. **Data:** 数据管理菜单，有关数据变量定义、数据格式选定、观察对象的选择、排序、定义多重响应集、加权处理、数据文件的转换、连接、汇总、产生正交表等。
5. **Transform:** 数据转换处理菜单，有关数值的计算、重新编码、建立时间系列、缺失值替代、产生随机数等。
6. **Analyze:** 统计分析菜单，集中了一系列在应用中所要用到的统计方法。
7. **Graphs:** 作图菜单，有关统计图的制作等。
8. **Utilities:** 实用程序，包括变量、OMS 标志、数据文件注释、定义变量集、使用变量集、运行手稿文件等。
9. **Add-ons:** 附加内容，提供包括 Amos、数据挖掘、抽样功效、数据录入、文本分析等应用程序，提供统计咨询、统计培训服务、可扩展的编程能力以及从三本统计手册中查找相关统计方法的说明等。
10. **Window:** 窗口管理菜单，有关窗口的排列、选择、显示等。
11. **Help:** 帮助菜单，有关帮助文件的调用、查寻、显示等。

点击主菜单选项即可激活菜单，这时会弹出下拉式子菜单，用户可根据自己的需求再点击子菜单的选项，来完成特定的功能。

从以上各主菜单主要从事的任务可见，SPSS16.0 的基本功能包括数据管理、数据计算、统计分析、图表分析、输出管理以及可扩展的编程能力等。

1.1.2 SPSS 的主要统计分析功能

SPSS 的统计分析功能主要集中在 Analyze 的主菜单中。单击 Analyze 弹出 Analyze 的子菜单，见图 1-2。各子菜单对应的统计功能见表 1-1。

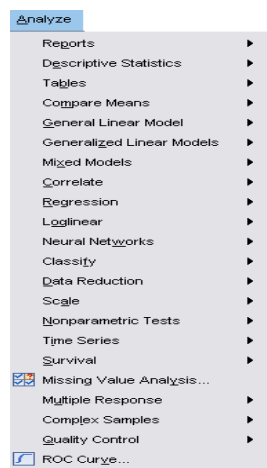


图 1-2 统计分析菜单

表 1-1 各子菜单对应的统计功能

子菜单名称	中文名称	统计功能
Reports	统计报表	制表、汇总
Descriptive Statistics	描述统计	计算描述统计量
Tables	表格	卡方检验
Compare Means	均数比较	比较均数差异
General Linear Model	一般线性模型	方差分析
Generalized Linear Models	广义线性模型	预测
Mixed Models	混合线性模型	预测
Correlate	相关分析	关联分析
Regression	回归分析	预测
Loglinear	对数线性分析	预测
Neural Networks	神经网络	预测
Classify	聚类分析	分类
Data Reduction	数据简化分析	降维
Scale	尺度分析	量表有效性与可靠性分析

(续表)

子菜单名称	中文名称	统计功能
Nonparametric Tests	非参数检验	分布一致性检验
Time Series	时间系列分析	趋势分析
Survival	生存分析	估计、预测
Missing Value Analysis	缺失值分析	缺失值处理
Multiple Response	多重响应	问卷整理、频数分布
Complex Samples	复杂抽样	抽样
Quality Control	质量控制	控制
ROC Curve	受试者工作特征曲线	控制

1.1.3 SPSS 的菜单与应用统计分析工作的主要步骤的对应关系

纵观现有 SPSS 中提供的应用程序的统计分析功能,虽然它有将别学科的数据处理方法逐渐引入的趋势,但整体而言,它基本上还是立足于数理统计的基本原理的。

数理统计是一门以概率论为基础,把带有随机性的数据作为研究对象,其任务是如何以有效的方法收集、整理和分析这些数据,并利用所得数据对所观察的现象做出推断或预测,以为决策提供依据。

数理统计的内容大体包括三个部分,一是数据资料的采集,二是数据资料的描述,三是统计推断。

数据资料的采集主要包括抽样方法(如何从总体中抽取样本)和实验设计(如何用最经济、最少次数的实验来获取与做大量实验等效结果的方法)等内容。

数据资料的描述是统计学的基础,侧重于研究对各个领域中的客观事物进行数字的计量、概括和表述方法,即主要研究对实验或调查中得到的大量数据资料如何进行科学整理,计算派生指标,制作统计图、表,找出这些数据的分布特征,计算出一些具有代表性的统计数字,用这些概括性的数字对总体特征进行简要的描述。

统计推断,它研究如何根据样本的特征推断总体的特征,即在描述统计的基础上,利用样本数据传递的信息,通过局部的研究来对总体的情形加以推断,并标明这种推断成立的可能性的。推断统计是当前统计学研究的主流。

上述这些内容将贯穿于统计的分析工作中。对于一般的统计分析工作,大致要经历如下的步骤,见图 1-3。

上面框图中所提到的整理数据资料,就是对所收集到的原始数据,进行审核、归纳、分组,并正确地按照统计分析方法的要求把有效数据输入计算机的相关统计软件(如 EXCEL、SPSS、SAS、Foxpro 等)中,形成数据文件,以便进行统计计算和分析。详见

本书的第 2 章。

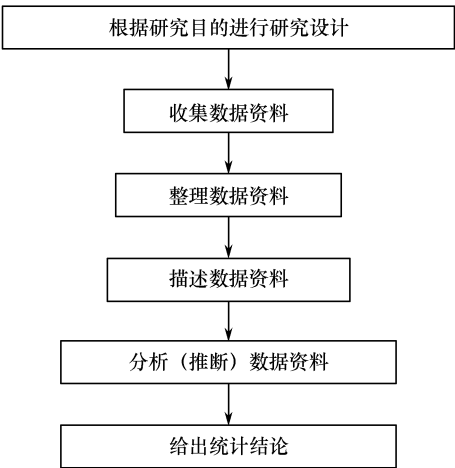


图 1-3 一般统计分析工作步骤

在 SPSS 中，各子菜单的过程里提供了一般的统计分析工作步骤所需要的大部分方法，不过它的分类并非按一般的统计分析工作步骤来分的，在有些过程里，它们是集数据资料的整理、描述和推断功能于一身的，因此，适当加以区分和归类，对我们快速进入工作状态是有益的。

我们将 SPSS16.0 菜单中的内容按其功能与一般的统计分析工作步骤中的研究设计、整理数据资料、描述数据资料和统计推断四步建立起初步联系。其对应关系见表 1-2。

表 1-2 一般统计分析工作步骤与 SPSS 菜单和子菜单的对应关系

内容	主菜单	子菜单	备注
实验设计	Data	Orthogonal Design	正交设计
	Analyze	Complex Samples*	复杂抽样设计
整理数据资料	File	New	新建数据文件
		Open	打开文件，读取数据文件
		Read Text Data	读取文本数据文件
		Save	保存当前数据文件
		Save As	另存当前数据文件
		Print	打印
	Data	Define Variable Properties	定义变量属性
		Define Dates	定义日期
		Merge Files	合并文件

(续表)

内容	主菜单	子菜单	备注
	Data	Split File	拆分数据文件
		Sort Case	样品排序
		Aggregate*	汇总数据
		Weight Cases	加权处理
	Transform	Compute variable*	计算派生变量
		Recode into Different Variable	重新编码
		Automatic Recode	自动编码
		Rank Cases	样品赋秩
		Date and Time Wizard	建立日期和时间变量
		Create Time Series	建立时间系列
		Replace missing Values	替换缺失值
		Random Number Generators	产生随机数
	Graphs		制作统计图
描述数据资料	Scale*		问卷调查的可靠性分析
	Analyze	Reports	统计报表
		Table*	制作表格、多重响应集
		Multiple Response	多重响应集
		Missing value Analysis	缺失值分析
	Analyze	Descriptive Statistics*	计算集中趋势和离中趋势统计量
		Survival*	生存分析
统计推断	Analyze	除上述提及子菜单以外的其他子菜单	差异或关联推断、预测、控制

注：表中带*号者表示除其具有本项的主要功能外，同时还具有其他项的功能。

1.2 研究设计

研究设计是科学研究工作中的重要一环，它做得好与坏将直接关系到取得数据资料的代表性、有效性和可靠性，影响到统计结论的可信度。一般地研究设计是根据研究课题来制定的，它主要包括实验设计和调查设计。研究设计主要解决：研究指标的选择与设计；研究（被试）对象的确定、抽样方式的设计及其相应研究精度的确定和抽样方式下的样本容量的最低限的估计；对应的统计方法的确定和科研经费的预算等。

1.2.1 研究指标的选择与设计

1. 研究指标

不同的研究领域对指标的定义是各不相同的。在新华字典中对一般的“指标”定义为事先规定的应达到的目标。也指检查、统计工作中实际达到的标准。

显然，在这里所说的“指标”同字典中的指标是同字不同意的。

在不同的学科中，对指标的定义也不尽相同。

在统计理论与统计设计上所使用的统计指标是指反映总体现象数量特征的概念。它包括三个构成要素：指标名称，计量单位，测量或计算方法。

在研究中，反映个别现象的叫做个体指标，反映总体现象的叫做综合指标。由于统计的任务主要是反映整体的数量特征，因此，在实际研究中，需要对一个研究的整体作出分解，选择和设计反映各部分的个体指标，并对不是常用的新设指标，要对其范围和计算方法做出合乎实际的具体规定，这些具体规定通常称为“统计指标口径”。目的是在其他研究者采用同样的指标作研究时，可使用相同的标准，以便进行对比分析。

2. 设计研究指标应当遵循的原则

(1) 以理论假设为指导

一项研究的目的在于检验研究提出的理论假设，因此，所选择和设计的研究指标必须能全面反映理论假设的内容，不能主观、随意地去罗列一大批与研究理论假设毫无关系的指标。指标设计工作常采用演绎方法，由理论假设到研究目标，从研究目标到研究变量，再由研究变量到研究指标。因此，设计指标时，应首先明确理论构思与假设，然后确定研究目标，弄清所涉及的各种研究变量，最后根据这些变量的客观要求，来制定收集实际数据与资料的指标。并由此构成一个有内在逻辑联系的、完整的研究指标体系。

(2) 可行性

所设计的研究指标，既不是在数量上越多越好，也不是在操作上越复杂越好。只要能全面、完整地反映理论假设与研究变量的主要维度即可，应尽可能删去一切不必要的多余指标，注意使研究指标简化，特别要考虑所设计指标的可行性，在操作上越简单的指标越具有可行性。

(3) 可重复性

为使不同的研究人员在作相同的研究中获得相同的结论，对研究中所设计的指标要使用操作定义明确地表述它们，保证它能够被观察、测量和重复操作。

1.2.2 研究对象的确定

任何一项研究都要有具体的研究对象，研究对象可以是人、是物，也可以是文献记

载或其他文字资料等，数量可以是一个，几个，也可以是成千上万个。

研究对象的确定不是随便的。它由以下几个方面决定：

首先，研究对象的确定取决于研究的目的，由研究的目的确定什么可作为研究的被试对象。例如，在“对中小學生实施分层递进教学，提高教学质量”的研究中，其研究对象就是中小学在校学生。一般地，对所要达到研究目的所关注的那个群体可以作为研究的对象。例如，在教育、教学科学研究中，由于我们往往需要关注影响教育、教学质量的因素，因此，此时的研究对象一般来说主要是指学生、学校、教师等这样的一些群体。

其次，无论什么样的研究对象，都必须要保证通过对它进行科学研究能够得出可靠的结论。如果研究对象是文献记载或其他文字资料，它必须是公开发行的或在国家的档案馆等地方可查证且得到各方认可的；如果研究对象为人、物或动物时，这就需要在确定研究对象时，能确信所选定的研究对象是可以通过某些确定的条件来判定其是否属于同类，同时还能确信选定的这些研究对象是符合研究的一些具体要求后参与到试验或调查中来的。在相同情况下所做的重复研究应有相同或相似的结论。

第三，确定研究对象应考虑其现实性和可能性，即根据研究者所具备的条件能否对研究对象进行研究。有些史料确实存在，但由于尚未到解密期等原因，暂时无法接触到，因此，这些资料不能作为研究对象。

第四，确定研究对象时还应考虑研究效率和效益，尽可能在比较短的时间内，以比较少的投入取得较大的研究成果。

1.2.3 抽样设计

在调查研究和实验研究中，通常情况下是通过从研究所关注的那个集合中，抽取部分研究对象来加以研究，用以此获取的信息来对其所隶属的整个群体做出估计、推断。

所谓抽样设计是指抽样方案设计者在充分利用抽样框的辅助信息和各种概率抽样方法的特点的基础上，制定一个切实可行和精度满足要求且最经济高效的如何抽样的方案。它不但要包括从抽样框中抽出哪些单位，还包括调查失败时的补救措施和获取调查数据后如何计算主要信息量的公式。

抽样应满足随机化原则。抽样要尽可能做到随机，即使全体研究对象中的每个个体的入样的机会均等，这样可以避免研究者的主观倾向或人为因素造成抽样偏差。

一般地，常见的抽样调查包括普查、概率抽样调查和典型调查三类。

(1) 普查

所谓普查是对研究对象的全体进行的全面调查。如，全国人口普查，全民身体素质普查等，它费时长，而且还要投入大量的人、财、物力。一般不会经常进行，通常普查的周期在 10 年以上。

(2) 概率抽样调查

概率抽样调查是在非全面调查中运用概率统计理论指导抽样调查的方法。它根据研究对象总体中一些已知的信息,充分利用它们,并据此设计合适的抽样方案,从而获得好的有代表性的样本。它与普查相比,可以节省大量的人力、物力、财力,还能大大缩短调查的时间。实施按概率统计原理设计的抽样方案得到的样本还可以对每一个特征指标进行估计,并给出估计的误差,它是目前许多领域获取调查信息时最公认和最常用的一种抽样调查方法。

(3) 典型调查

典型调查是一种完全依靠先验经验的抽样调查。由于它抽取的样本含量较少,因此样本对总体代表性的好坏,很大程度上取决于设计取样方案的调查者掌握的先验信息。因而,它往往要以普查和概率抽样调查为基础确定典型样本。

上述三类抽样调查的配合使用,可以获得正确而时效性很强的总体信息。

1.2.4 样本容量的确定

样本容量取多少合适是研究设计中必须慎重考虑的一个环节。虽然,样本容量越大,其代表性越好,但是,随着样本容量的增加,势必会增大研究中的人力、物力和财力的投入,因此,综合起来考虑未必就是样本容量越大越好。

样本容量的大小取决于以下一些因素。

1. 研究的类型和范围

当研究是定量研究,研究范围较广,样本容量可适当大些;反之,研究是定性研究,研究范围较窄,样本容量可适当小些。

2. 研究的精度

研究的精度越高,要求的样本容量越多。

3. 允许误差限

用样本统计量推断总体参数时,允许误差限越小,要求的样本容量越多。

4. 总体的同质性

当总体的同质性较好时,样本容量可以取得少些,如同一个人的血液的同质性较高,所以化验时只需几滴血即可;当总体的变异性较大时,样本容量应取得大一些。

5. 总体容量

一般地,总体容量越大,所取的样本容量也应较大些。

6. 测量工具的可靠程度

当测量工具的可靠程度较低时,测量的误差就比较大,需要的样本容量也应大一些,反之,可适当减少些样本容量。比如,在心理学测试中,一般学习能力、成就的测量工具的可靠性程度好一些,此时,样本容量可适当少些;而人格特质、自我概念和态度等

方面的测量工具的可靠性程度差一些,此时,样本容量可适当大些。

7. 研究经费

研究经费宽松时可以适当增加一些样本容量,但研究经费紧张时,需要严格控制成本,此时,只能量体裁衣了,能满足样本容量的最低要求即可。

8. 分析类别

当研究指标较多,它们之间的关系较复杂时,需要大一些的样本容量,反之,在一些单指标的研究中,样本容量可适当少些。

在第 3 章几种常用的概率抽样方法中,专门论述有关上述涉及的概念和样本容量具体确定的方法。

1.2.5 实验设计中用到的一些基本术语

1. **试验指标** (因变量): 试验中需要考察的指标。如研究不同锻炼方法对减肥的效果,不同运动强度对百米跑运动成绩的影响,则体重、百米跑成绩等就是试验指标,简称指标。

2. **试验因素**: 试验中需加以考察的各种因素称为试验因素。例如,要研究不同的运动量、运动强度和运动持续时间对运动成绩的影响,则运动量、运动强度和运动持续时间等就称为试验因素,简称因素。

3. **单因素方差分析**: 当考察的试验因素只有一个时,称单因素方差分析。如,当参与试验的被试对象的各方面条件都基本相似时,探讨同一个老师的几种不同的教法对学生学习成绩的影响,就是单因素方差分析。因为此时影响学习成绩的因素只有一个教法因素。

4. **多因素方差分析**: 当考察试验的因素有两个或两个以上时,称其为多因素方差分析。例如,为了找出适合某个专项运动的运动强度、运动量、运动持续时间的较佳组合,我们需要考察不同的运动强度、运动量和运动持续时间等因素对专项运动成绩的影响,因此,它是一个多因素方差分析。

5. **水平**: 每个因素所处的不同状态称水平。例如,将运动强度分为:大、中、小三种不同的状态,则称将运动强度分为三个不同的水平。

6. **处理**: 在试验中,所有因素各取定一个水平组成一个试验条件,称这些试验条件为处理,又称单元。例如研究问题中的因素有性别,取值为 0、1;有年龄,分三个水平 1 (10 岁)、2 (11 岁)、3 (12 岁)。两个变量的组合共可形成六个处理 ($C_2^1 C_3^1 = 2 \times 3 = 6$): [1,1]、[1,2]、[1,3]、[2,1]、[2,2]、[2,3],代表两种性别与三种年龄的六种组合。在方差分析中,比较各处理下,因变量均值之间的差异。

7. **试验单元**: 试验中被安排在一个处理的试验单位称为试验单元。

8. **因素的主效应和因素间的交互效应**: 在多因素方差分析中,由于参与实验的因素

至少有两个, 如因素 A 与因素 B, 因素 A、B 对试验结果的影响称为因素的主效应, 而因素 A、B 的联合作用对试验结果的影响称为因素的交互效应。

9. **全面试验**: 如果每一个可能的处理都做试验, 称为全面试验。

10. **部分试验**: 只从所有处理中挑选一部分处理进行试验观察, 这样的试验就称为部分试验。

11. 实验设计的原则

1935 年, R.A.Fisher 在他出版的《试验设计法》中, 试验设计应遵循三个原则, 即重复、随机排列和局部控制。

重复是指在一个试验中同一处理设置两个以上的试验单位。这样可以估计试验误差, 也可以降低试验误差。

随机排列是指试验中每个处理都有相等的机会安排在任何一个试验单位上。它和重复原则结合在一起使用, 就能提供无偏的试验误差估计。随机要贯穿在整个试验过程的始终。

局部控制就是分范围、分区域控制非试验因素, 使各处理所受的影响趋于最大限度地一致。

1.2.6 常用的实验设计

差异问题主要关注两个或更多组或条件之间是否有显著性差异。在问到分组比较或差异问题时, 能把独立变量和设计归结为组间或组内。

1. **单组设计**: 从考察的总体中随机抽取一个样本含量满足要求的被试对象, 从它们身上测定某个或某些观察指标的数值, 同给定的分布或分布参数的常量进行差异检验。这种实验设计方案还称为标准对照。如果测试指标只有一个, 叫做一元单组设计, 如果测试指标有多个, 叫做多元单组设计。

2. **成组设计**: 实验中, 仅涉及一个具有两水平的实验因素。当实验因素的水平与被试对象的分组无关时, 可将全部的被试对象随机地分成两组, 分别接受不同的处理, 所以也称为完全随机成组设计。而当实验因素的水平与被试对象的分组有关时, 只能在特定的被试对象所在的总体中随机各抽取一个样本, 此时称为组内随机成组设计。由于在研究中的每个被试对象在且只能在一个条件或一个组中出现, 所以它属典型的组间设计。当试验指标只有一个时, 叫做一元成组设计, 也称实验对照。当试验指标有多个时, 叫做多元成组设计。

3. **配对设计**: (1) 自身配对设计, 在同一个被试对象身上, 两次测得同一个定量测试指标的值, 这些值成对出现。它也称为组内设计, 即研究中的各个被试对象, 经受或经历独立变量的所有的条件或水平。(2) 同源配对设计, 用来自母体相同的两个个体, 进行配对, 在同一个定量测试指标上各测试一次, 从而得到成对数据。(3) 条件相近配

对设计,用各种条件相近者组成配对,在同一个定量测试指标上各测试一次,从而得到成对数据。当试验指标只有一个时,叫做一元配对设计,当试验指标为多个时,叫做多元配对设计。

4. **单因素多水平设计:** 实验中,涉及一个具有 k 个水平 ($k>2$) 的实验因素。当实验因素的水平与被试对象的分组无关时,可将全部的被试对象随机地分成 k 组,分别接受不同的处理,所以也称为完全随机设计。而当实验因素的水平与被试对象的分组有关时,只能在特定的被试对象所在的总体中随机各抽取一个样本,此时称为组内随机设计。由于在研究中的每个被试对象在且只能在一个条件或一个组中出现,所以它属典型的组间设计。当试验指标只有一个时,叫做单因素设计。也称实验对照。当试验指标有多个时,叫做多元单因素设计。

5. **随机区组设计:** 将全部被试对象按区组因素分成若干组,每个区组内的被试对象间互相接近,再将每个区组内的被试对象随机地分到每个处理组中。当试验指标只有 1 个时,称为随机区组设计。当试验指标有多个时,称为多元随机区组设计。

6. **双因素无重复实验设计:** 将两个实验因素中的一个放置在横向上,一个放置在纵向上,设横向因素有 R 个水平,而纵向因素有 C 个水平,将全部 $R \times C$ 个原始条件基本相似的被试对象随机地分到各个处理的单元中去,每个单元只有一个被试对象。

7. **析因实验设计:** 利用纵向和横向两个方向来排列全部实验因素及其水平,使实验因素之间的全部水平组合都能以纵横交叉的形式呈现出来,各种水平组合条件下至少做两次或两次以上的独立重复试验。当实验因素与被试对象无关时,也可将原始条件基本相似的被试对象随机地分到各个处理的单元中去,使每个单元有 n 个 ($n \geq 2$) 被试对象。

8. **混合析因设计:** 根据某个或某些实验因素被试对象完全随机地分成几个独立的组(有一个组间变量),接受处理后,再在几个不同的时间点上从同一个被试对象上重复获得指标的观察值(一个组内独立变量),则称它为混合设计。混合设计在有实验前测试和实验后测试的实验研究中是通用的。

9. **正交试验设计 (Orthogonal experimental design):** 是研究多因素多水平的一种高效率、快速、经济的试验设计方法。日本著名的统计学家田口玄一将正交试验选择的水平组合列成表格,称为正交表。正交试验设计依托正交表,根据正交性从全面试验中挑选出部分有代表性的点进行试验,这些有代表性的点具备了“均匀分散,齐整可比”的特点,正交试验设计是分式析因设计的主要方法。例如做一个三因素三水平的实验,按全面实验要求,须进行 27 次组合实验,且尚未考虑每一组合的重复数。若按 $L_9(3)^3$ 正交表安排实验,只需做 9 次,大大减少了工作量。因而正交实验设计在很多领域的研究中已经得到广泛应用。

1.2.7 在研究设计中的统计分析预案

在研究设计中, 需要给出对收集到的数据资料使用何种统计方法进行统计分析的预案。

如何选择统计方法, 这是科研人员最关心的问题。有些人错误地认为, 对调查或试验中获取的数据资料, 用 SPSS 处理后就能得到科学的统计结论, 结果事实与之正相反。许多研究的成果之所以得不到大家的认可, 许多花费了很大财力和精力的研究中, 总存在着明显的瑕疵和缺陷, 达不到理想的效果, 这往往与误用统计方法有关, 有些则是由于实验设计存在先天缺陷所造成, 还有一些是由于只用了一些适用的统计方法, 但数据资料中的许多信息还没有选择到更合适的统计方法将其完整地表述出来所造成。究其原因, 从本质上而言, 恐怕与其对统计原理和方法的不理解有关。

所以, 单就统计方法的选择而言, 建立起对统计基本原理的理解, 掌握常用的实验设计方法以及相应的统计分析方法, 这是正确选择统计方法的必要的前提和基础。本书中也会涉及到一些这方面的理论和知识, 但本书关注的是如何引导你正确选择 SPSS 中的最适宜的方法, 如何阅读统计结果并合理地解释之, 所以更详细的统计理论方法的内容, 则要在专门的统计书籍中获得, 因此, 阅读一些专门的统计书籍对于理解本章中正确选择统计方法肯定是有益的。

许多刚从事研究的人员尚未建立起良好的科研习惯, 不按正常的科研程序办事, 总是喜欢在实验研究和调查研究结束后, 等到要进行数据资料分析时, 才想起该用什么样的统计方法来处理手中的数据资料。实际上, 这绝不是个好习惯。有时免不了要走弯路, 甚至还会造成无可挽回的损失。例如, 有位研究人员要做实验前、后数据资料的各指标间的相关和差异性检验, 由于实验前没有考虑好应用什么样的统计方法处理实验结果以及这些方法对数据的要求, 因而在收集数据资料时, 未按配对资料的要求去对测试指标的值做完整的记录, 只记录了具体测试指标的值。到数据资料分析处理时, 已变成了一堆与研究目的毫不相关的无用数据, 只得重新安排实验, 从头再来。

因此, 统计方法的选择绝不应该在数据资料分析时进行, 而是要在研究课题确定后所做的研究设计或调查设计中就应该有预案。

一般来说, 统计方法的选择是个复杂的系统, 涉及多方面的因素, 它主要与研究目的有关, 显然还要涉及研究设计、调查设计和数据资料的类型以及分布类型等。因此需要将它们综合起来一起考虑, 才能最终确认所需要的合适的统计方法。

例如, 为了探讨不同缺氧方式影响肺泡表面活性物质代谢规律, 将 36 只家兔随机分成 4 组, 每组 9 只。一组为对照组, 一组为急性缺氧组, 一组为间断缺氧 5d 组, 还有一组为间断缺氧 15d 组, 实验观测肺泡支气管灌洗液中 5 种磷脂的相对含量, 即溶血磷脂酰碱、磷脂酰碱、磷脂酰甘油、神经鞘磷脂、磷脂酰乙醇胺, 则这个实验设计是属于完

全随机单因素 4 水平 5 元变量的实验设计。由于测试指标值是计量数据资料，所以，在研究设计中，统计分析的预案是：实测数据资料后，可以根据数据资料的分布类型和协方差阵或方差的齐性，来选择相应的统计分析方法，具体过程参见图 1-4、图 1-5。

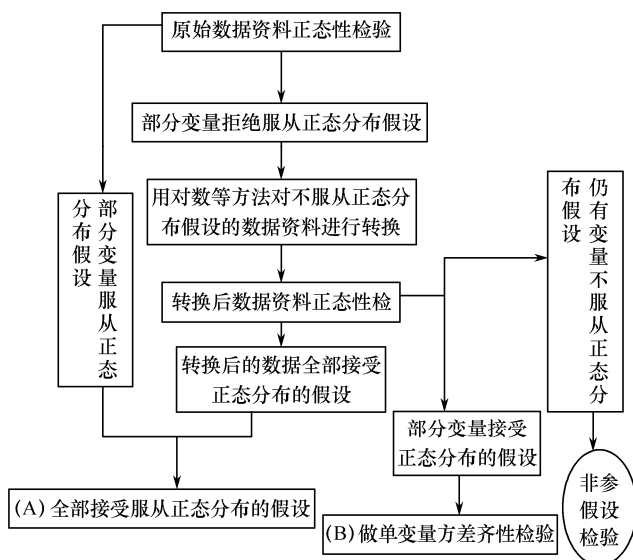


图 1-4 统计方法选择

图 1-4 中的 (A) 和 (B) 后接图 1-5 中 (A) 和 (B) 后续部分。

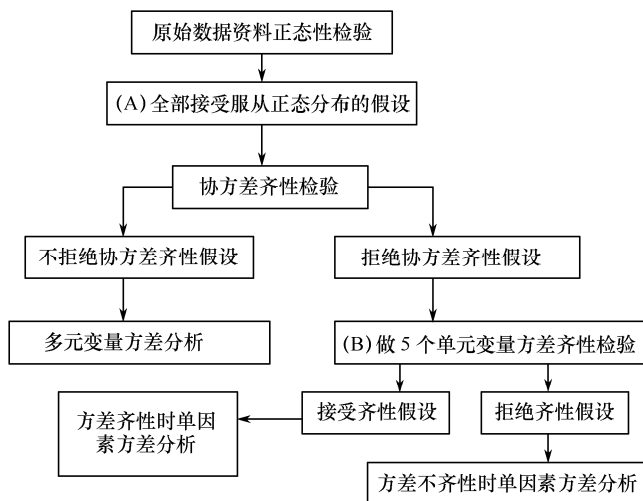


图 1-5 统计方法选择

1.2.8 在研究设计中需要用到 SPSS 中的子菜单、过程和程序

1. 计算抽样的样本含量

在计算抽样所需的样本含量时需要用到的 SPSS 中的子菜单、过程和程序见表 1-3。

表 1-3 计算抽样的样本含量时需要用到的 SPSS 中的子菜单、过程和程序

步骤	子菜单	调用过程	程序	对话框中选项	选择	功能
1	Analyze	Descriptive Statistics	Descriptives	Options	在 Options 对话框中选择 Variance	计算方差
2	Transform		Compute Variable	在 Function group 中选择 All	在 Functions and special Variables: 中选择 IDF.Normal	设置公式 计算所需 样本含量

2. 抽样设计

在 SPSS 中进行抽样设计时，可以按以下方式进入选择一个样本的设计工具：

依次单击 Analyze→Complex Samples→Select a Sample，在打开的抽样工具中按提示要求，选择应答和输入必要的参数，进行抽样设计。

3. 产生正交表和正交设计方案

依次单击 Analyze→Data→Orthogonal Design→Generate，在展开的对话框中可以设置因素名、水平数和最少试验次数，从而在当前工作的数据编辑窗口中产生所设定的正交表。

依次单击 Analyze→Data→Orthogonal Design→Display，展开 Display Design 对话框，在该对话框中通过设定打印的因素名，从而可在输出窗口中生成正交试验设计的方案。

1.3 整理数据资料

1.3.1 在 SPSS 中建立数据文件

1. 新建数据文件

在当前数据文件窗口，按 File→New 顺序，可打开一个新的数据文件编辑窗口。

2. 定义变量

在编辑窗口下，单击 Variable view 按钮，在 Variable view 编辑窗口中，可以对变量进行定义。

在数据编辑窗口中已有数据的情况下，也可按 Data→Define Variable Properties 顺序，在展开的 Define Variable Properties 对话框中，选择需要定义的变量，在其后的 Define Variable Properties 的工具中对变量进行定义或修改。

3. 插入变量、插入样品

按 Edit→Insert Variable 顺序, 可以在选中的列前插入一个新变量。

按 Edit→Insert Vacases 顺序, 可以在选中的行前插入一个新样品。

4. 替换数据

按 Edit→Replace 顺序, 在展开的 Find and Replace 对话框中, 定义寻找和替换的数据, 对数据文件中的某些数据进行替换。

5. 排序

按 Data→Sort Cases 顺序, 在展开的 Sort Cases 对话框中, 顺序选择需要排序的变量组, 对数据文件进行按样品的数值大小进行排序。

按 Data→Sort Variables 顺序, 在展开的 Sort Variables 对话框中, 选择用来排序的变量的名或标签、值等变量的内容, 对数据文件的变量进行重新排列。

1.3.2 在 SPSS 中读取数据文件

1. 读取 SPSS 数据文件

按 File→Open→Data 顺序, 可选择打开一个已经存在的数据文件。

2. 读取其他数据库的数据文件

如果需要将在 dBASE、Excel 或 MS Access Database 中生成的数据文件, 转换成 SPSS 数据文件, 则可按下述步骤进行:

按 File→Open Database→New Query 顺序, 展开 Database Wizard 对话框, 在其工具中, 选定数据文件的类型, 再选择数据文件名, 可在 SPSS 数据编辑窗口读取其他数据库的数据文件。

3. 读取文本数据文件

按 File→Read Text Data 顺序, 展开 Open Data 对话框, 选择文本文件存放的途径和文件名, 可在 SPSS 数据编辑窗口读取文本数据文件。

1.3.3 在 SPSS 中合并数据文件

1. 将其他数据文件中的样品数据增加到工作数据文件的记录后面

当数据文件中的变量名相同, 需从另一个数据文件向当前工作数据文件合并数据时, 可按下述步骤进行:

按 Data→Merge Files→Add Cases 顺序, 展开 Add Cases 对话框, 选择需要合并的文件名, 即可。

2. 将其他数据文件中的变量及数据添加到工作数据文件中

当两个合并的数据文件中有一个共同的关键变量, 而其他变量名不同, 需从一个数据文件向当前工作数据文件添加不同变量的数据时, 可按下述步骤进行:

按 Data→Merge Files→Add Variables 顺序, 展开 Add Variables 对话框, 选择需要合并的文件名, 指定共同的关键变量即可。

1.4 分析数据资料

总的说来, 在研究过程中, 研究人员一般主要关注三类研究问题, 即如何对统计资料作整体的描述, 比较研究中不同总体在相同测试指标上是否有差异性, 从同一批被试对象中测得的不同指标间是否存在关联性等。对应于这三类研究问题, 需要正确选择相应的 SPSS 统计过程、程序和统计量。

1.4.1 研究目的是对数据资料作一般性描述

当研究的目的中, 有需要用到对收集到的原始资料作整体描述时, 采用描述统计的方法是正确的选择。

描述统计是统计分析工作的基础, 侧重于研究对特定领域中的客观事物进行数字的计量、概括和表述方法, 即主要研究对实验或调查中得到的大量数据资料在用科学的方法进行整理的基础上, 找出这些数据的分布特征, 计算出一些具有代表性的统计数字, 用这些概括性的数字对总体特征进行简要的描述。

1.4.1.1 描述统计过程概述

在 SPSS 中, 涉及描述统计的内容较多, 主要集中在主菜单 Analyze 下拉式菜单中的 Reports、Descriptive Statistics、Tables 三个子菜单中和主菜单 Graphs 下拉式菜单中的 Chart builder、Legacy Dialogs 二个子菜单中。

Reports 子菜单中对定性和定量资料进行按行、列等不同方式进行分层、分组统计, 产生记录报表, 同时可计算一些简单的描述统计量。

Descriptive Statistics 子菜单中包括了一系列描述统计的分析过程, 可以进行常用的频数分布表分析, 对定量资料包括正态分布在内的各种常见分布的检验和探索分析, 制作 P-P 图和 Q-Q 图, 以及在二维列表中进行卡方检验等。

Tables 子菜单中, 主要针对分类资料产生各种所需的行*列表、频数表, 并可产生复杂的多层/嵌套表及建立多重应答集。

Graphs 子菜单中, 提供了非常全面的统计图库, 比其他统计软件有更强大的制图功能, 利用豪华的操作界面可以制作出非常精美的统计图。

有关 SPSS 中描述统计的过程及其基本功能见表 1-4。虽然有些方法上都有计算某些统计量的功能, 似乎有些重复, 但仔细推敲, 实质上各种方法侧重点各有不同, 这反过来给使用者提供了更大的自由选择的空间。

表 1-4 描述统计的过程与基本功能

菜单	子菜单	过程	基本功能描述
Analyze	Reports	OLAP Cubes	生成多层表，计算常用统计量
		Case Summaries	对变量按指定格式进行记录列表，并计算相应的统计量
		Report Summaries in Rows	对输出表格作精细定义，按行输出相应的统计量，为纯文本格式
		Report Summaries in Columns	对输出表格作精细定义，按列输出相应的统计量，为纯文本格式
	Descriptive Statistics	Frequencies	产生频数分布表
		Descriptives	进行一般描述统计
		Explore	计算常用统计量，并进行正态性和方差齐性检验
		Crosstabs	对定性资料进行常规统计计算与检验， χ^2 检验
		Ratio	对两个尺度变量计算相对比
		P-P Plots	估计分布类型并制作 P-P 图
		Q-Q Plots	估计分布类型并制作 Q-Q 图
	Tables	Custom Tables	自制表格并计算常用统计量和作最基本的统计分析
		Multiple Response Tables	专门为多选题数据设计制表过程
Graphs	Chart builder	无	用预先设定的图库或单独的制图程序制图
	Legacy Dialogs	Bar	制作条形统计图
		3-D Bar	制作 3 维条形统计图
		Line	制作线图
		Area	制作面积图
		Pie	制作饼图
		High-Low	制作高低图
		Boxplot	制作箱图
		Error Bar	制作误差图
		Population Pyramid	制作金字塔图
		Scatter/Dot	制作散点图
		Histogram	制作直方图
		Interactive	制作交互式图

1.4.1.2 描述统计过程的适用条件

在 SPSS 中提供的这些描述统计的方法，并不是任意一个测试指标变量都能用的，同样还要受到其他条件的限制。一般而言，研究指标的数据资料类型成为选择具体描述统计方法的首要条件。关于统计资料类型的划分，将在第 2 章 2.1 节中进行讨论。主要的资料类型是定量资料和定性资料。

1. 不适用于定性资料的过程

在 SPSS 提供的描述统计的方法中，以下过程不适用于定性资料，见表 1-5。

表 1-5 不适用于定性资料的描述统计过程

菜单	子菜单	过程	适用资料类型
Analyze	Reports	OLAP Cubes	定量资料
	Descriptive Statistics	Descriptives	定量资料
		Explore	定量资料
		Ratio	定量资料
		P-P Plots	定量资料
		Q-Q Plots	定量资料

2. 不适用于定量资料的过程

在 SPSS 提供的描述统计的方法中，以下过程不适用于定量资料，见表 1-6。

表 1-6 不适用于定量资料的描述统计过程

菜单	子菜单	过程	适用资料类型
Analyze	Descriptive Statistics	Crosstabs	定性资料
	Tables	Multiple Response Tables	定性资料

3. 既适合于定量也适合于定性资料的过程

在 SPSS 提供的描述统计的方法中，以下过程既适合于定量也适合于定性资料的过程，见表 1-7。

表 1-7 适合于定性和定量资料的过程

菜单	子菜单	过程	适用资料类型
Analyze	Reports	Case Summaries	定性和定量资料
		Report Summaries in Rows	定性和定量资料
		Report Summaries in Columns	定性和定量资料
	Descriptive Statistics	Frequencies	定性和定量资料
	Tables	Custom Tables	定性和定量资料

(续表)

菜单	子菜单	过程	适用资料类型
Graphs	Chart buider	无	定性和定量资料
	Legacy Dialogs	Bar	定性和定量资料
		3-D Bar	定性和定量资料
		Line	定性和定量资料
		Area	定性和定量资料
		Pie	定性和定量资料
		High-Low	定性和定量资料
		Boxplot	定性和定量资料
		Error Bar	定性和定量资料
		Population Pyramid	定性和定量资料
		Scatter/Dot	定性和定量资料
		Histogram	定性和定量资料
		Interactive	定性和定量资料

1.4.1.3 描述统计过程中可分析的定量资料的分布类型

对定量资料而言，如何正确选择合适的统计量来描述定量资料的集中趋势的关键是要了解该测试指标的总体的分布类型。

在 SPSS 有关描述统计的过程中，涉及分析定量资料分布类型的过程，见表 1-8。

表 1-8 定量资料分布类型的分析过程

Analyze 下的子菜单	过程名称	可分析的分布类型及相关的检验
Descriptive Statistics	Explore	正态分布、方差齐性检验
	P-P Plots	正态分布、学生氏分布、对数正态分布、指数分布、卡方分布、伽马分布、拉普拉斯分布、逻辑斯谛分布、半正态分布、帕累托分布、均匀分布、威布尔分布
	Q-Q Plots	

1.4.1.4 定量资料描述统计分析过程中的常用统计量

定量资料分析中，可能要用到许多描述集中趋势和离中趋势及整体趋势描述的统计量，这些统计量在许多过程中可能重复出现，但有些也只出现在某个特定的过程中，为便于研究人员在具体的描述统计分析工作中，更有针对性地选择描述统计过程，将含有描述统计常用统计量的过程汇总在一起，具体参见表 1-9 和表 1-10。

需要注意的是，大部分的统计量在其过程的主对话框中可以找到并能选择，但也有一些表中提到的统计量只有在其执行后的输出结果中才能看到。

表 1-9 定量资料描述统计分析过程中的常用统计量 (I)

统计量	主菜单	Analyze			
	子菜单	Reports			
	过程	OLAP Cubes	Case Summaries	Report Summaries in Rows	Report Summaries in Columns
集中趋势统计量	算术平均数	√	√	√	√
	几何平均数	√	√	×	×
	调和均数	√	√	×	×
	中位数	√	√	×	×
	众数	×	×	×	×
	总和	√	√	√	√
	总和百分比	√	√	×	×
离中趋势统计量	最大值	√	√	√	√
	最小值	√	√	√	√
	两极差	√	√	×	×
	四分位差	×	×	×	×
	百分位数	×	×	×	×
	方差	√	√	√	√
	标准差	√	√	√	√
	标准误差	√	√	×	×
	变异系数	×	×	×	×
	百分比	×	×	√	√
分布统计量	峰度	√	√	√	√
	偏度	√	√	√	√
	正态分布检验	×	×	×	×
	方差齐性检验	×	×	×	×
总数	样本含量	√	√	√	√
	占总数百分比	√	√	×	×

表 1-10 定量资料描述统计分析过程中的常用统计量 (II)

统计量	主菜单	Analyze				
	子菜单	Descriptive Statistics				Tables
	过程	Frequencies	Descriptives	Explore	Ratio	Custom Tables
集中趋势统计量	算术平均数	√	√	√	√	√
	几何平均数	×	×	×	×	×
	调和均数	×	×	×	×	×
	中位数	√	×	√	√	√
	众数	√	×	×	×	√
	总和	√	√	×	×	√
	总和百分比	×	×	×	×	√
离中趋势统计量	最大值	√	√	√	√	√
	最小值	√	√	√	√	√
	两极差	√	√	√	√	√
	四分位差	√	×	×	×	×
	百分位数	√	×	√	×	√
	方差	√	√	√	×	√
	标准差	√	√	√	√	√
	标准误差	√	√	×	×	√
	变异系数	×	×	×	√	×
	百分比	×	×	×	×	√
分布统计量	峰度	√	√	√	×	×
	偏度	√	√	√	×	×
	正态分布检验	×	×	√	×	×
	方差齐性检验	×	×	√	×	×
总数	样本含量	√	√	√	√	√
	占总数百分比	√	√	√	√	√

1.4.1.5 对定量资料描述统计分析时的统计量选择

对定量资料用描述统计计算统计量时, 需要考虑测试指标的分布类型, 在测试指标的总服从正态分布或对称分布时, 描述数据资料的集中趋势统计量可用算术平均数, 而测试指标的值是等比级数的数据资料时, 也即它服从对数正态分布时, 描述它的集中趋势统计量就要选用几何均数, 而当测试指标的分布是偏态分布时, 描述数据资料的集

中趋势统计量可用中位数或众数。

1.4.1.6 描述统计分析时常用的统计图表

无论是定量资料还是定性资料，描述统计中最直观的方法是制作统计图、表。常用的表格有：行汇总表、列汇总表、交叉表、自定义表等，而常用的统计图有：直方图、P-P图、Q-Q图、圆图、散点图、条形图、线图、面积图、箱图和误差条图等。

1.4.1.7 选择描述统计方法的一般步骤

综上所述,当研究目的需要对研究问题作描述统计时,可根据资料的类型从表 1-5 至表 1-7 中选择适宜的统计的过程,如果是定量资料,则从表 1-9、表 1-10 中,选择相应的方法对其分布进行分析和检验,再根据数据资料的分布类型,按 1.4.1.5 中的要求,确定选择合适的统计量进行计算,根据表 1-9、表 1-10 中统计量集中所在的过程,选择相对集中的过程进行分析,根据研究目的,从表 1-7 中选择相应的图、表制作过程,制作相应的统计图、表。

1.4.2 研究目的是对不同总体在相同指标上是否有差异进行推断

推断统计是统计分析过程中的重要内容之一，它研究如何根据样本的特征推断总体的特征，即在描述统计的基础上，利用样本数据传递的信息，通过局部的研究来对总体的情形加以推断，并标明这种推断成立的可能性的太小。

推断统计的方法，分布在 SPSS 的 Analyze 主菜单的下拉式子菜单中的绝大部分过程中。正确选择推断统计方法是每个研究者都以求的。

1.4.2.1 差异性检验中合适的推断统计方法的选择

1. 关于一元单组设计时差异性推断统计方法的选择

(1) 测试指标为尺度类型且服从正态分布, 同一个已知正态总体均数作差异性检验时, 可选择 Analyze → Compare Means → One-Sample T Test...来实现。

(2) 测试指标为名义类型且服从二项分布, 同一个已知率作差异性检验时, 可选择 Analyze → Nonparametric Tests → Binomial... 来实现。

(3) 测试指标为有序类型且服从多项分布, 同多项已知率 (或已知期望频数) 作差异性检验时, 可选择 Analyze → Nonparametric Tests → Chi-Square... 来实现。

(4) 测试指标为尺度类型, 样本含量大于 100 时, 与参数未知的正态分布、均匀分布、泊松分布、指数分布作分布一致性检验时, 可选择 Analyze → Nonparametric Tests → One-Sample Kolmogorov-Smirnov... 来实现。

(5) 测试指标为尺度类型, 样本含量大于 3 小于 5000 时, 与参数未知的正态分布作分布一致性检验时, 可选择 Analyze → Descriptive Statistics → Explore..., 在 Explore 对

话框中，单击 **Plots** 按钮，在 **Plots** 对话框中，选择 **Normality plots with tests** 选项来实现。

2. 关于单元单因素设计时的差异性推断统计方法的选择

当有两个变量，其中一个是因变量时，此时选择差异性检验的适宜推断统计方法时，首先要根据因变量的测度类型，是尺度还是名义来做第一步区分；第二步，如果是尺度类型，根据样本含量的大小，可用 1.4.2.1 的 1 中的（5）或（4）的方法来判定它是正态分布还是非正态分布；第三步，结合因素变量的设计方法是单组、成组还是配对来加以综合考虑。具体选择方法参见表 1-11。

表 1-11 为两个变量的差异性检验选择一个合适的推断统计

因变量 测度类型	比较	单因素 2 水平		单因素多水平	
		成组比较	配对比较	完全随机设计	组内随机设计
尺度因变量服从正态分布	平均数	Analyze → Compare Means → Independent-Sample T Test...	Analyze → Compare Means → Paired-Sample T Test...	Analyze → Compare Means → One Way ANOVA	Analyze → General Linear Model → Repeated-Measures
尺度因变量不服从正态分布	平均秩	Analyze → Nonparametric Tests → 2 Independent Samples Mann-Whitney U	Analyze → Nonparametric Tests → 2 Related Samples Wilcoxon	Analyze → Nonparametric Tests → K Independent Samples Kruskal-WallisH	Analyze → Nonparametric Tests → K Related Samples Friedman
因变量是名义或两分变量	计数	Analyze → Descriptive Statistics → Crosstabs Chi-Square	Analyze → Nonparametric Tests → 2 Related Samples McNemar	Analyze → Descriptive Statistics → Crosstabs Chi-Square	Analyze → Nonparametric Tests → K Related Samples Cochran Q

3. 关于多元单组设计时多元正态总体均值等于常数向量检验时的推断统计方法

用各自的样本观察值减去相比较的各自已知的正态总体均值后所形成的新的样本数据值，选择 **Analyze → General Linear Model → Multivariate** 来实现作差异性的推断统计。

4. 关于多元单因素设计时多元正态总体的差异性推断统计方法的选择

可以选用多因变量线性模型的方差分析，即选择 **Analyze → General Linear Model → Multivariate** 来实现作差异性的推断统计。

5. 关于单元或多元多因素设计时的差异性推断统计方法的选择

根据因变量的个数、测度水平和分布及试验设计的类型，来选择合适的推断统计方法，见表 1-12。

表 1-12 为三个或三个以上变量的差异性检验选择一个合适的推断统计

因变量	两个或多个因素（独立变量）		
	全部组间	全部组内	混合
一个正态/尺度因变量	Analyze → General Linear Model → Univariate	Analyze → General Linear Model → Repeated Measures	Analyze → General Linear Model → Repeated Measures
多个正态/尺度因变量	Analyze → General Linear Model → Multivariate	Analyze → General Linear Model → Repeated Measures	Analyze → General Linear Model → Repeated Measures
有序因变量	Analyze → Generalized Linear Models → Generalized Linear Models	Analyze → Generalized Linear Models → Generalized Estimating Equations	Analyze → Generalized Linear Models → Generalized Estimating Equations
两分因变量	Analyze → Generalized Linear Models → Generalized Linear Models; Analyze → Loglinear → Logit	Analyze → Generalized Linear Models → Generalized Estimating Equations	Analyze → Generalized Linear Models → Generalized Estimating Equations

1.4.3 研究目的是检查两个或多个变量之间的关联或相关

1. 为两个变量的相关或关联问题的假设选择一个恰当的推断统计
根据变量的测度水平、分布类型来选择合适的推断统计方法，见表 1-13。

表 1-13 为两个变量的相关和联合问题的假设选择一个恰当的推断统计方法

两个变量的测度水平	发生联系	相同或相关的被试对象的两个变量或得分
尺度变量且服从正态分布	得分	Analyze → Correlate → Bivariate(Pearson) Analyze → Correlate → Linear
两个变量或是有序变量或不服从正态分布	秩	Analyze → Correlate → Bivariate (Kendall's Tau-b 或 Spearman)
一个是正态尺度变量，另一个是名义变量		Analyze → General Linear Model → Univariate 在 Option 中选择 Estimates of effect size 看输出结果中的 η^2
两个变量是名义或两分变量	计数	Analyze → Descriptive Statistics → Crosstable 在 Statistics 中选择 Phi and Cramer's V

2. 为预测来自几个独立变量中单因变量选择恰当的多元统计方法
根据因变量的测度水平和自变量的分布类型和测度类型，来选择合适的多元统计方法，见表 1-14。

表 1-14 为预测来自几个独立变量中单因变量选择恰当的多元统计方法

单因变量 ↓	几个自变量或预测变量			
	所有正态/尺度	一些正态、一些或所有二分（2 类）	一些或所有正态（分类超过两类）	有至少一个随机和/或嵌套变量的正态和/或二分
正态/尺度 （连续的）	Analyze → Regression → Linear	Analyze → Regression → Linear; 或 Analyze → General Linear Model → Univariate	Analyze → General Linear Model → Univariate	Analyze → Mixed Models → Linear
二分	Analyze → Classify → Discriminant	Analyze → Regression → Binary Logistic	Analyze → Regression → Binary Logistic	Analyze → Generalized Linear Models → Generalized Estimating Equations

1.4.4 研究目的是缩减指标

在实际的科学研究中，有时反映一个整体的特征，往往需要用到很多的指标来同时进行刻画，由于指标一多，指标之间的关系就变得比较复杂，尤其在用到多元回归、判别分析等的多元统计分析时，由于变量之间的高度相关所产生的共线性问题，势必会影响统计分析的结果，因此，将变量归类，利用类间变量相对独立的特点，通过组合新变量或挑选各类的典型指标来减少变量之间的相关，就变得很有必要。此外，当研究指标较多时，也很难从主次关系的角度给出它们对整体的重要性，因此，通过适当的多元统计方法，将它们进行线性组合，变成少数几个综合指标，同时又能充分反映大量原始指标的信息，从带信息量的角度来区分综合指标的重要性，这也是非常有意义的。

1. 变量分类和求典型指标

按 Analyze → Classify → Hierarchical Cluster 顺序，在展开的对话框中，如果要做指标聚类，选择 Variables，如果要做样品聚类选择 Case。

对聚成一类的变量，通过它们两两之间的相关系数，利用公式

$$\bar{R}_j^2 = \frac{\sum r^2}{m_j - 1}$$

在按 Transform → Compute Variable 顺序展开的 Compute Variable 工具中，计算每个变量的相关指数。

2. 对尺度变量进行主成分分析、因子分析或对应分析

按 Analyze → Data Reduction → Factory 做主成分或因子分析，或按 Analyze → Data Reduction → Correspondence Analysis，对变量和样品做对应分析。

第2章 数据资料的整理与描述

所谓数据资料的**整理**就是对所收集到的原始数据，进行审核、归纳、分组，并正确地按照统计分析方法的要求把有效数据输入计算机的相关统计软件中，形成数据文件，对数据资料的异常值、缺失值情况和分布类型进行初步的判定，以便后续进行进一步的统计计算和分析，并在此基础上制作相应的统计图表。

而数据资料的**描述**是将数据资料的信息通过少数代表集中趋势和离中趋势的统计指标反映出来，便于后续的统计推断。

2.1 SPSS 数据文件的制作

本节的目的是建立 SPSS 数据文件。有关 SPSS 的安装、启动、数据编辑窗口的介绍、文件的存、取方式等基础知识方面的详细介绍，可参阅由卢纹岱主编的《SPSS 统计分析》第 1 章的相关内容。

俗话说得好，巧妇难为无米之炊，SPSS 也一样，当其数据编辑窗口的数据显示区（有行和列组成的二维平面表格）（见图 2-1）中没有任何数据，即没有数据文件时，除查看（View）菜单中的全部过程和文件（File）菜单中新建（New）、打开（Open）、打开数据库（Open Database）、读取文本数据（Read Text Data）、重新命名数据文件（Rename Dataset）等与向数据编辑窗口中录入、导入数据文件和数据文件改名等少数过程及其他菜单中的极少数过程能够运行外，SPSS 中的大部分功能，尤其是数据（Data）、转换（Transform）、统计分析（Analyze）、图形（Graphs）等菜单中的各个过程运行后都会出现如图 2-2 所示的对话框。这说明，调用 SPSS 各种程序之前，其数据编辑窗口中必须要有数据文件。

数据存放在数据编辑窗口的单元格（Cell）（见图 2.1）中，在单元格中可输入文字、字母、数字等信息，一个单元格中只能存放一个数据。单元格的顶框是变量名（Var），最左侧边框是样品号（Case）又称记录号。

在统计学中，将参与调查或实验研究中的每个研究对象称为被试对象又称样品或个体，从被试对象身上得到的调查指标或试验指标的结果称变量，由于从每个被试对象身上测得的同一测试指标的值极有可能是各不相同的，因此，变量是一个在特定区间中可以取各种各样值的量，这是变量的可变性，此外，在一个具体被试对象身上测定一个具体指标得到的值就是该变量的实测值。例如，用百分制测试某班学生的数学考试成绩，

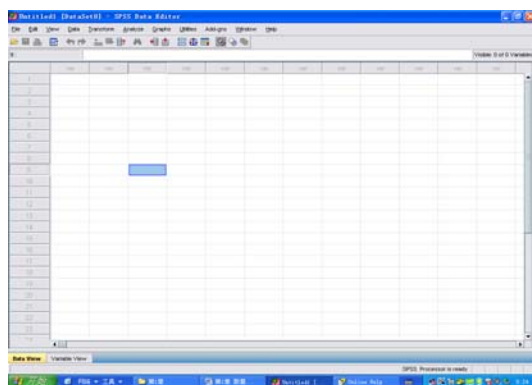


图 2-1 数据编辑窗口

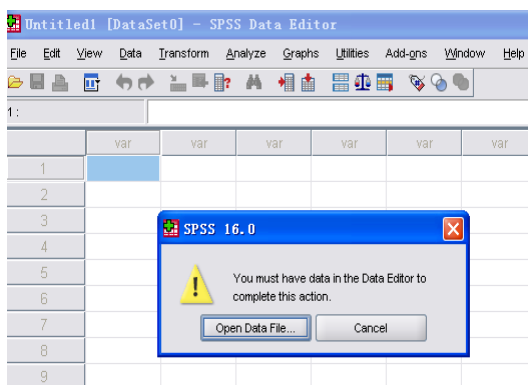


图 2-2 数据编辑窗口中必须有数据才能计算

则数学考试成绩是个统计指标也就是变量，它可在 0~100 分中取值，某个学生得分 90，这是变量的一个实测值。被试对象在 SPSS 中用样品来描述，第三个被试对象称为样品 3。因此从第三个被试对象上测得的第 4 个观察指标的值就应被录入第三个样品和第四个变量交叉形成的单元格中。

SPSS 的数据文件由变量、样品及非空的变量值组成。它可在 SPSS 数据编辑窗口中录入、导入或修改而成，也可以在（按 File→Open→Data 顺序打开的）Open Data 对话框中，选择其他格式的数据库文件转换而成。有关这方面的详细介绍，可参阅由卢纹岱主编的《SPSS 统计分析》第 2 章的相关内容，及本章 2.3 节中的内容。

2.1.1 统计资料的类型与变量类型、测度类型的对应关系

在 SPSS 中建立数据文件时，首先要涉及的是定义变量，而定义变量的关键是如何去指定它的类型和测度类型。变量的测度类型和变量类型同以后要遇到的各种统计分析过程直接有关，这一点已在第 1 章中作过阐述。变量的类型和测度类型的定义取决于调查指标和测试指标实测后所获得的统计资料的类型。

例 2.1 在参与某个反映城市居民收入和支出现状的调查研究的被试者中，随机抽出一位章先生的一些个人资料见表 2-1。

表 2-1 章先生的个人资料

姓名	性别	年龄	婚姻状况	家庭人口	汽车	家庭总收入（元）	学历	学位	身高（cm）	体重（kg）	爱好
章××	男	40	已婚	3	1	12058	研究生	博士	175	69	文学

从表 2-1 中易见，章先生的资料有文字和数字两部分组成。它可以说是统计工作中得到的数据资料类型的一个概貌。

2.1.1.1 统计资料类型

在章先生的个人资料中，我们观察到有两种资料类型：

一种是以数量形式表现的**定量资料**，如年龄、家庭人口、家庭总收入、身高、体重等，另一种是以属性或属性的程度来描述的**定性资料**，如性别、婚姻状况、爱好等。

定量资料根据取值的连续程度可分为：在一个给定范围内可以取任何一个值的连续型的**计量数据**（如身高、体重）和在整数范围内取值的离散型的**计数数据**（如家庭人口、每户的汽车数）。

定性资料根据属性和程度可分为：用数字代表事物属性分类的**名义数据**（如用 1、0 分别代表性别男、女，用数字 1、2、3、4 分别代表婚姻状况的未婚、有配偶、离婚、丧偶等）和用数字代表事物属性不同程度分类的**有序数据**（如用 1、2、3、4、5 分别代表学位：博士后、博士、硕士、学士、无学位等）。

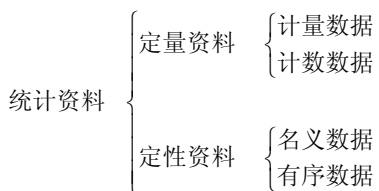


图 2-3 统计资料分类

因而可用图 2-3 来归纳上述的统计资料的分类。

显而易见，在以后 SPSS 统计分析过程中遇到的数据资料的类型也就是这四种类型中的一种或二种或多种的组合。

数据资料的类型将决定定义变量的类型和测度类型，从而在某种程度上限制了其使用的统计模型和方法。有时，根据研究的需要，计量资料也可以通过制作区间频数分布的方式，将其转化成计数资料或有序数据。其方法在第 3 章 3.1 节中讨论。从而可以从另外的角度用与计量资料不同的统计理论和方法来研究数据资料中隐藏的内在统计规律性。

2.1.1.2 SPSS 中的变量类型

1. 变量与常量的区别

常量是指一个不变的量，例如， $x \equiv 6$ ，则 x 就是常量。而变量必须至少能取两个不同的值。如有 1=男、0=女两个值都存在的**性别**就成为一个变量，但如果我们研究的样本中，性别都为男性，此时我们用性别描述的量就不再是变量，而是一个常量。

2. 变量类型

在 SPSS 数据编辑器的 Variable View 选项卡中，单击变量类型 (Type) 列的单元格，在弹出的 Variable Type 对话框图 2-4 中，可见 SPSS 中的变量类型可以分为数值型和字符型两种。其中数值型变量有 7 中不同的显示方式，分别是：标准数值型 (Numeric)、整数部分自右向左每 3 位用逗号作分隔符的数值型 (Comma)、同 Comma 正好相反，小数

点用逗号表示,整数部分自右向左,每3位用圆点作分隔符的数值型(Dot)、科学记数法数值型(Scientific Notation)、带美元符号的数值型(Dollar)、自定义数值型(Custom Currency)、日期数值型(Date)。而在Data菜单下的Define Variable Properties(定义变量属性)过程的Type选择项中,还有一种Percent,见图2-5,它是以百分比方式显示的数值型。

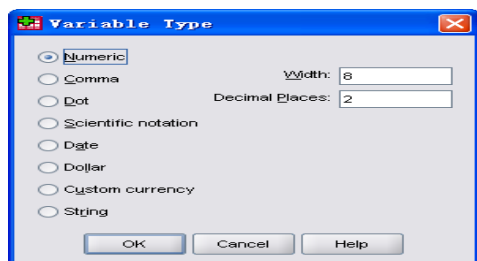


图 2-4 定义变量类型

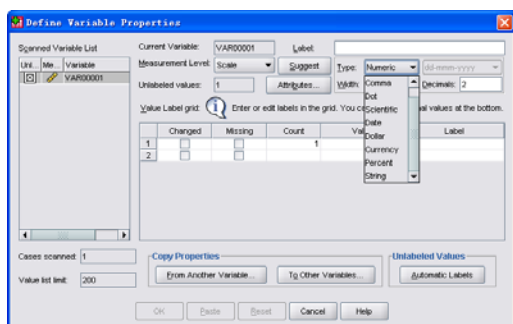


图 2-5 变量的类型

3. 变量类型的定义与修改

新建变量类型的定义,可在数据编辑器的Variable View选项卡中,单击变量类型(Type)列的单元格,在弹出的Variable Type对话框中选择(见图2-4)。

修改一个已经存在的变量类型,除可用定义新建变量类型相同的方法外,还可在Data菜单下选择Define Variable Properties(定义变量属性)过程,在展开的对话框中,见图2-6,选择变量名列表中需要修改的变量,将其拖曳到变量浏览(Variables to Scan)框中,系统默认扫描不同变量值数最大为200,如果不同的变量值数超过200,则选中Limit number of values displayed to后框中的数字进行修改即可。如果对观察的记录(也称样品)有限制,可选中Limit number of cases scanned to选项,并在其后框中输入需要限制浏览的记录数。

单击Continue按钮,展开定义变量属性对话框(图2-5)。在浏览变量名列表中选择需要浏览的变量,则在Type框中可选择变量类型。

上述限制浏览样品数和限制显示的变量值数两项选择项联用,意味着要浏览给定的最大记录数中,观察给定最大不同变量值的综合分布情况。见图2-5中Value Label Grid下的显示内容。

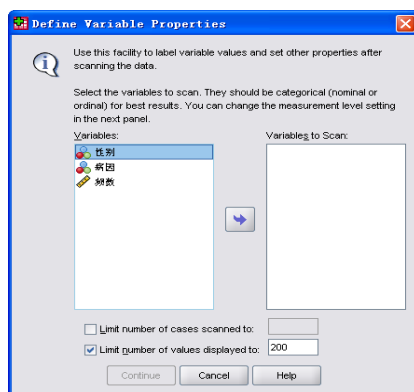


图 2-6 选择定义变量属性浏览变量

如果选择了自定义 (Custom currency) 类型, 则在 Edit 菜单下的 Option 过程的对话框中的 Currency 选项卡中进行定义。具体做法参阅由卢纹岱主编的《SPSS 统计分析 (第 4 版)》第 1 章的相应内容。

2.1.1.3 SPSS 中的变量测度类型

在 SPSS 中, 系统定义的测度类型有三种类型, 它们分别是: 尺度测度 (Scale)、有序测度 (Ordinal) 和名义测度 (Nominal)。

在数据编辑器的 Variable View 选项卡中, 单击测度类型 (Measure) 下的单元格, 可以定义该行变量的测度类型。同样, 也可在 Data 菜单下选择 Define Variable Properties (定义变量属性) 过程, 按 2.1.2 中的方法, 在 Measurement Level 下拉菜单中 (见图 2-5) 选择测度类型。

2.1.1.4 统计资料的类型与变量类型和测度类型间的对应关系

一般而言, SPSS 中的有序测度 (Ordinal) 和名义测度 (Nominal) 属分类变量的测度类型, 它对应于调查研究中遇到的定性资料, 而尺度测度是专门为定量资料设计的。设立名义数据和有序数据的目的是为了区分类别, 而非本身之间的计算, 所以一般可将其设定为字符型, 但当它是字符型时, 其值标签将在输出的表格中不起作用。如果我们清楚分类变量一般不直接参与运算, 作为减少存储数据文件时所占用的空间, 也可将其变量类型定义为数值型。存储一个数值型数据比一个字符型数据要少占 8 个字符的空间。定量资料肯定是要用来计算分析的, 所以其变量类型为数值型。故数据资料的类型与变量类型和测度类型间的对应关系可参见表 2-2。

表 2-2 数据资料的类型与变量类型和测度类型间的对应关系表

统计资料类型	数据类型	变量类型	测度类型
定量资料	计量数据	数值型	尺度
	计数数据		
定性资料	名义数据	字符型或数值型	名义
	有序数据	字符型或数值型	有序

2.1.2 建立调查问卷的数据文件

虽然实验设计中涉及实验条件的控制和实验指标的测定的要求是十分严格的, 但仅从在 SPSS 中整理数据, 建立数据文件的角度而言, 与问卷调查相比, 无论是从涉及的面, 还是从题型的多变性上, 实验设计要相对简单得多。

例 2.2 下面是从某个研究者在其研究设计中所做的问卷调查设计部分选取的一些问题。

北京市体校学生调查问卷

编号：_____（调查者不用填写此项）

姓名：_____、性别：_____、年龄：_____周岁、参与训练的年龄：_____周岁。

问题 1、请在下面的选项中选择你所在的学校：_____。

- (1) 一体校 (2) 二体校 (3) 三体校 (4) 首体院竞技体校
(5) 五体校

问题 2、你的运动等级是：_____。

- (1) 国家健将 (2) 国家一级 (3) 国家二级、(4) 国家三级 (5) 无级

问题 3、请在下列选项中选择你毕业时的意向：_____。（可多选）

- (1) 进专业队 (2) 升学 (3) 就业 (4) 其他

问题 4、请在以下影响运动成绩的因素选项中，按你认为的重要性的的大小做出排序选择：_____。

- (1) 运动强度 (2) 运动量 (3) 运动持续时间

问题 5、如果你能在大学中继续从事你的运动专项，在下面的选项中，你最想去的地方是：_____。（可多选）

- (1) 大学高水平运动队 (2) 体育系 (3) 运动系 (4) 军警校
(5) 职业学院 (6) 其他

问题 6、如果有以下单位可供你毕业后选择，你想去的单位是：_____。（答案可多选）

- (1) 中学 (2) 小学 (3) 业余体校 (4) 机关学校 (5) 体育场馆
(6) 企业工会 (7) 国家机关 (8) 其他选择

问题 7、请在下述各项上，给自己的水平和能力作如实的评价（在选中项上打√）

	很好	较好	一般	不好	很不好
专业知识水平					
计算机水平					
英语水平					
语言表达能力					
示范能力					
教学训练能力					

问题 8、你家兄弟姐妹的人数为：_____。

本例的目标是在 SPSS 中将其建成数据文件。

我们已经知道，数据文件由变量名、样品号和变量值组成，尽管在 SPSS 中，对已经建成的数据文件还可以修改，但这主要是针对变量值而言的，如果修改变量的类型，很可能会导致原先对变量定义的某些信息的丢失。例如，将数值型名义测度变量更改成字

符型名义测度变量,则原定义的值标签内容丢失。因此,在实际调查问卷数据较多的情况下,为避免不必要的返工,及多人能同时开展工作,在 SPSS 中录入数据资料之前,必须首先建立起统一规范的数据文件格式,养成这个良好的习惯非常重要,而要做到这一步的关键是必须先要理清调查问卷中的题型与建立变量数之间的关系。

2.1.2.1 调查问卷中的题型分类

从例 2.2 中,根据提问方式,可以总结出,一般调查问卷中,问题的题型基本可分为七种类型:

1. 名义单选题,即在众多不分程度的可选答案中,只能选择其中一种答案。如,问题 1. “请在下面的选项中选择你所在的学校: _____。(1) 一体校 (2) 二体校 (3) 三体校 (4) 首体院竞技体校 (5) 五体校”,这种题型就是名义单选题。

2. 有序单选题,即在众多有程度可分的可选答案中,只能选择其中一种答案。如,问题 2. “你的运动等级是: _____。(1) 国家健将 (2) 国家一级 (3) 国家二级 (4) 国家三级 (5) 无级”,这种题型就是有序单选题。

3. 多选题,即在众多的可选答案中,可以选择其中一个或多个答案。如,问题 5. “如果你能在大学中继续从事你的运动专项,在下面的选项中,你最想上的地方是:(可多选)(1) 大学高水平运动队 (2) 体育系 (3) 运动系 (4) 军警校 (5) 职业学院 (6) 其他”,这种题型就是多选题。

4. 排序题,即在众多的可选答案中,选择时要按被试者认为的重要性的将其排成一个有序的答案数列。如,问题 4. “请在以下影响运动成绩的因素选项中,按你认为的重要性的做出排序选择: _____。(1) 运动强度 (2) 运动量 (3) 运动持续时间”。这种题型就是排序题。

5. 单空题,需要被测者根据实际情况自己填写,但一个问题中,只有一个空需要填写。如,问题 8. “你家兄弟姐妹的人数为: _____。”

6. 多空题,即在一个问题中,有多个空需要被测者根据实际情况自己填写。如,“姓名: _____、性别: _____、年龄: _____(周岁)、参与训练的年龄: _____(周岁)。”等。

7. 多重有序(名义)单选题,有多个有序(名义)单选题组合在一起形成的综合题。如,问题 7 的题型。

2.1.2.2 调查问卷中的题型与变量设计

虽然在一个单元格中,通过字符串的方式,可以将多选题或排序题的结果存放在一起,但是这样做后统计起来将十分不便,如果不加任何技术处理,SPSS 中的现成的统计分析程序是不能处理的。因此,为便于后续的计算分析,原则上对应于同一个样品的每一个变量中只能出现一个观察值。

所以,在 SPSS 中输入上述题型的结果做数据文件时,要为每个单空题和每个单选题

的结果建立一个相应的变量，这同实验研究中，要为每个测试指标建立一个单独变量是一样的，而要为多选题、多选题和排序题要建立多个变量来分别存放选择的结果，即有多少个空或多少个可供选择的选项，就要设与之等同数量个数的变量。这相当于将一个多选题变成多个具有二分变量（其可供选择的选项只有两个，即 1（是）、0（否））的单选题。而将排序题变成多个有序单选题。而将多选题变成多个单选题。对于多重有序（名义）单选题，也要设多个变量将其变成一个具体的单选题来处理。

因此，对于例 2.2，应建立的变量的数量见表 2-3。

表 2-3 例 2.2 每题建立变量对应表

	被试者基本情况	问题 1	问题 2	问题 3	问题 4	问题 5	问题 6	问题 7	问题 8
建立变量数	5	1	1	4	3	6	8	6	1
合计	35								

而对于变量的类型和测度类型可参照 2.1.4 中表 2-2 的对应关系来设定。

2.1.2.3 变量命名

只要在 SPSS 数据编辑器的任意一个单元格中输入信息，该单元格所在列及左侧各列的表格顶端显示系统提供的默认变量名。从左向右依次为：VAR00001、VAR00002、VAR00003、…，直到录入数据的单元格所在列为止。

例如，在第 5 行和第 4 列交叉的单元格中录入数据 6 回车，则在 SPSS 数据编辑器中显示的内容见图 2-7。录入的数字 6 自动变成系统默认的带有 2 位小数有尺度测度的标准数值型。由 6 所形成的左上矩形区域中的其他所有单元格中，系统默认其数值为缺失值。所自动生成的 4 个变量，均为标准数值型。同样，如果在第 5 行和第 4 列交叉的单元格中录入字符 A 回车，则系统自动判定第 4 列的变量为字符型名义测度，同上，由 A 所形成的左上矩形区域中的其他所有单元格中，系统默认其值为缺失值。左侧 3 列的变量系统默认认为是标准数值型。

由此可见，为避免不必要的系统默认缺失值的存在，录入资料应从第一个单元格开始。同时，上面的例子也让我们清楚，在 SPSS 中，变量名就是对变量的命名不是必需的。尤其对于那些临时用于探究性研究的数据文件而言，更是如此。

但在实际的研究工作中，建立起来的数据文件，并非只是在一次研究中使用，而是有一定的连续性，尤其在比较研究中，若干时间后我们还可能会用到以前的数据资料，这就需要使用变量名或标签来查看数据文件存放的资料。

	VAR00001	VAR00002	VAR00003	VAR00004
1
2
3
4
5	.	.	.	6.00
6

图 2-7 录入数字 6 后的显示图

此外,每次研究中,我们都会要用到 SPSS 统计处理的结果,这些统计处理的结果多数是以图、表的方式来表示的,对 SPSS 处理得到的原始图、表,原则上我们不应对它做修改,否则会让人联想你是不是也修改了表中计算得到的真实数据结果。如何让读者清楚地了解在图、表中,你处理的是什么变量,正确使用好变量名也是必需的。因此,可以说对变量的命名是实际工作的需要,有很重要的实际意义。

1. 变量命名的原则

在 SPSS12.0 以上的版本中,为变量命名要遵循以下几个原则:

- 首字符必须是字母或汉字(在 SPSS16.0 中,还可以是@、#或\$之一,但用户的自定义变量名中,#或\$不能作为首字符,在首字符后,它们连用在一起是符合规则的),其后可以是任何字母、数字、圆点或符号@、#、_、或\$;

- 变量名不能用圆点结束。也要避免用下划线结束;
- 变量名最多不能超过 64 个字符(在 SPSS12.0 及以下版本中最多可以是 8 个字符);
- 空格和特殊字符(如? , ', ! 和*) 不能使用;
- 每个变量名必须是唯一的,不允许重复;
- 变量名不能同 ALL、AND、BY、EQ、GE、GT、LE、LT、NE、NOT、OR、TO、WITH 等 SPSS 的保留字相同;

- 变量名可以用大小写混合的字母定义,并为显示目的原样保留,但系统对同一个字符的大小写认为是相同的字符;

- 在输出中,当较长的变量名需要重叠在多行中时,SPSS 试图在下画线、圆点以及小写字母向大写字母改变处分行。

上述 SPSS 中的变量命名的原则是我们在实际工作中应遵循的一个底线。在实际问题中,对变量命名还要考虑到不同的专业和涉及的领域,用词要尽量简洁、明了,要能让人一目了然地理解你所要表达的实际含义。

对一个具体的调查测定内容作简要描述时,可采用以下实际命名的原则和方法:

在 SPSS12.0 以下版本中使用时,首先,要考虑用约定俗成的简称对变量命名,以便能在 4 个汉字或 8 个字符内完成命名。如测试某样本的身高、体重值后,可分别用身高、体重来直接命名变量名。其次,要尽量使用专业术语来表示变量。例如,测试两臂侧伸时两手指尖的最大长度得到数据资料,要对该测试指标的变量命名,可用“臂展”来描述。

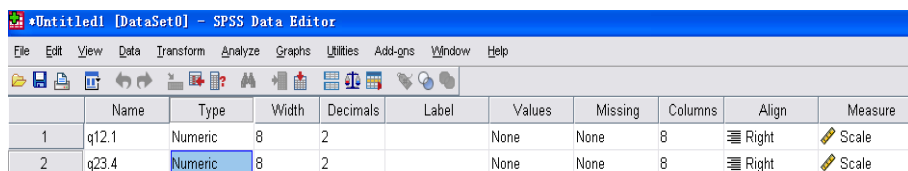
对于无法在 4 个汉字内命名的变量,也可用定性描述其测试属性的前几个汉字的汉字拼音的首字母命名变量,而将详细的描述体现在其 Label 中。详见 2.1.2.4。

对于问卷调查中的问题的命名,建议采用 Q(question 的第一个字母)后接题号的方式表示,如第 3 题可用 Q3 表示变量名,第 13 题可用 Q13 表示变量名,而有些题需要表示其中小题的,可在大题命名的基础上加“.小题号”表示其变量名,如 15 题 2 小题,可用 Q15.2 表示变量名,该题的具体内容可在 Label 中作说明。

即使在 SPSS12.0 及以上版本中使用时, 变量名也应尽量要短, 不应定义得太长, 否则输出的表格太庞大, 不美观。更重要的是它会造成不必要的纸张的浪费。

2. 定义变量的方法

在 SPSS 数据编辑窗口中, 单击 Variable View 选项卡, 进入变量命名和定义状态, 见图 2-8。



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	q12.1	Numeric	8	2		None	None	8	Right	Scale
2	q23.4	Numeric	8	2		None	None	8	Right	Scale

图 2-8 变量定义

变量定义包括: 变量名 (Name)、类型 (Type)、宽度 (Width)、小数位数 (Decimals)、标签 (Label)、值标签 (Values)、缺失值定义 (Missing)、列宽 (Columns)、对齐方式 (Align)、测度类型 (Measure) 等定义。

对具体的变量而言, 上述各项定义并不一定全部都要做一遍, 有些可以用系统提供的默认值即可, 有些也只要定义其中的一部分即可。

例如, 你所要定义的变量是定量资料, 如身高, 则除在变量名下输入身高以外, 其他项可直接采用系统提供的定义。

考虑到定性资料的变量定义后, 同以后选取的统计分析方法、输出表格直接有关, 所以定义变量时, 一般要重点关注变量名 (Name)、类型 (Type)、标签 (Label)、值标签 (Values)、测度类型 (Measure) 等的定义。

光标置于第一列的某个单元格中就可定义该行的变量名, 如果变量名为中文, 可选用系统默认的 Ctrl+Shift 组合键选择输入法 (当然也可使用你自定义的组合键来选择输入方法), 输入完毕后按回车, 在该单元格中出现键入的中文名。例如, 输入性别后回车, 则该变量被命名为性别。如果要输入英文变量名, 按 Ctrl+Space 组合键即可进入 ASCII 码输入法状态。Ctrl+Space 组合键是中英文输入法状态的转换键, 在 SPSS 中同样适用。

双击变量名单元格可进入修改变量名状态。

在变量名不能让我们很清晰理解它所表达的含义时, 可单击该行对应 Label 列所对应的单元格, 在此, 我们可输入对变量名的说明, 尽管它的最大字符数为 255 个, 但变量标签不是越长越好, 在注解清楚的前提下, 应越短越好。

基于我们对该变量在研究中所要起的作用, 及根据 2.1.4 中的对应关系, 在 Type 列中定义该变量的类型, 在 Measure 列中定义该变量的测度类型。对定性资料, 我们用分类变量名义或有序定义其测度类型。因而有必要对其值标签做出相应的定义。在 Values

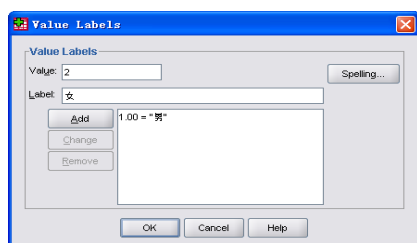


图 2-9 定义值标签

列中可以定义该变量的值标签。单击该单元格右侧的省略号，展开 Value Labels 对话框，见图 2-9。在 Value: 框中输入 1，在 Value Label 后框中输入男，单击 Add 按钮，在最下面的框中加入如图 2-9 所示的一个值标签。重复这样的过程，可将该变量中的所有的不同值定义其值标签。

2.1.2.4 建立数据文件实例分析

下面以例 2.2 为例来说明数据文件的建立。

1. 变量定义

根据资料类型、题型和变量命名的原则，对例 2.2 中涉及到有关问题定义以下的变量，见表 2-4。

表 2-4 对例 2.2 调查问卷建立数据文件时的变量定义

变量	变量名	类型	变量名标签	值标签	测度
1	编号	字符型			名义
2	姓名	字符型			名义
3	性别	数值型		1=男、0=女	名义
4	年龄	数值型			尺度
5	参训年龄	数值型			尺度
6	所在学校	数值型		1=一体校、2=二体校、3=三体校、4=首体院、5=五体校	名义
7	运动等级	数值型		1=国家健将、2=国家一级、3=国家二级、4=国家三级、5=无级	有序
8	进专业队	数值型		1=选中、0=未选	名义
9	升学	数值型		1=选中、0=未选	名义
10	就业	数值型		1=选中、0=未选	名义
11	其他	数值型		1=选中、0=未选	名义
12	运动强度	数值型		1=第一重要、2=第二重要、3=第三重要	有序
13	运动量	数值型		1=第一重要、2=第二重要、3=第三重要	有序
14	运动持续时间	数值型		1=第一重要、2=第二重要、3=第三重要	有序
15	Q5.1	数值型	上大学高水平运动队	1=是、0=不是	名义
16	Q5.2	数值型	上体育系	1=是、0=不是	名义

(续表)

变量	变量名	类型	变量名标签	值标签	测度
17	Q5.3	数值型	上运动系	1=是、0=不是	名义
18	Q5.4	数值型	上军警校	1=是、0=不是	名义
19	Q5.5	数值型	上职业学院	1=是、0=不是	名义
20	Q5.6	数值型	其他	1=是、0=不是	名义
21	Q6.1	数值型	中学	1=是、0=不是	名义
22	Q6.2	数值型	小学	1=是、0=不是	名义
23	Q6.3	数值型	业余体校	1=是、0=不是	名义
24	Q6.4	数值型	机关学校	1=是、0=不是	名义
25	Q6.5	数值型	体育场馆	1=是、0=不是	名义
26	Q6.6	数值型	企业工会	1=是、0=不是	名义
27	Q6.7	数值型	国家机关	1=是、0=不是	名义
28	Q6.8	数值型	其他选择	1=是、0=不是	名义
29	专业知识	数值型	专业知识水平	1=很好、2=较好、3=一般、4=不好、5=很不好	有序
30	计算机	数值型	计算机水平	1=很好、2=较好、3=一般、4=不好、5=很不好	有序
31	英语水平	数值型		1=很好、2=较好、3=一般、4=不好、5=很不好	有序
32	语言表达	数值型	语言表达能力	1=很好、2=较好、3=一般、4=不好、5=很不好	有序
33	示范能力	数值型		1=很好、2=较好、3=一般、4=不好、5=很不好	有序
34	教训能力	数值型	教学训练能力	1=很好、2=较好、3=一般、4=不好、5=很不好	有序
35	兄妹人数	数值型			尺度

2. 录入数据资料

在录入大批量的资料信息时，首先要给回收来的调查问卷编号，调查问卷的编号应同数据文件中的编号相一致。这样做的目的是有利于后面资料的审核工作。建议不要用数据编辑窗口中的样品号代替编号，原因是当对数据文件的某个变量作排序处理后，数据文件中样品顺序会发生改变，而数据编辑窗口显示区边框上的样品号是不变的。这就是为编号单设一个变量的原因，录入姓名的目的也是为审核录入数据资料的正确性。当然，如果将来作样品聚类，你也能很快知道哪些被调查者被聚到一起。但不管怎样，姓名变量在后续的统计计算分析中是用不上的。在数据文件资料审核通过后，可以删除该变量。

在录入资料信息时，不要将已习惯的单变量的工作经验移植到相关的多变量的工作中来。一张调查问卷中的数据应录入在同一行相应变量对应列的单元格中，不能错行、错列。

当建立了变量名下的单元格中含有有效样品的信息时，便得到了 SPSS 的数据文件。将其存盘后所形成的文件，称为 SPSS 磁盘数据文件。SPSS 磁盘数据文件的扩展名为 *.sav。

由例 2.2 建成的数据文件见 data02-01.sav。

2.1.3 将 EXCEL 中建立的数据文件变成 SPSS 中数据文件

例 2.3 在 EXCEL 工作簿“第 2 章例 2.3 相关与回归数据资料.xls”中的第一张工作

	A	B	C	D	E
1	13岁时身高	母亲身高	13岁时足长	13岁时骨龄	17岁时身高
2	164	160	23.4	-1.4	172
3	164	162	22.2	-1	177
4	162	158	25.2	-1.8	180
5	160	159	22	4	167
6	158	167	24.2	1	160
7	158	161	21.5	0	162
8	156	160	22.6	-1.8	172
9	154	170	23.1	-4	180
10	151	152	21.2	2	159
11	148	155	20.4	2	158

图 2-10 13 岁男孩身高及相关指标的资料

表中已建立了准备预测 13 岁男孩身高的数据文件，如图 2-10 所示。为下一步在 SPSS 中作多元回归分析做准备，先将其变成 SPSS 中的数据文件。

很多时候，最初的数据文件并不一定是直接在 SPSS 的数据编辑窗口的显示区中制成的。有的可能在 Lotus 中、有的可能在 dBASE 中、有的可能在 Text 中被做成数

据文件。本例只是众多可能中最常见的一种。虽然，建立数据文件时用的软件各不相同，但 SPSS 已充分考虑到与这些数据分析软件间的数据对接问题，因此从本例 EXCEL 的数据文件向 SPSS 数据文件转化的方法与途径中，也可清楚知道其他软件中所建的数据文件是如何转换的。

本例有两种方法可完成数据文件的转换工作：复制法和打开法。

2.1.3.1 复制法

这是一种大家比较习惯的做法，具体步骤如下：

1. 打开“第 2 章例 2.3 相关与回归数据资料.xls”，在工作表 1 中，将鼠标的光标移到 A2 单元格，按住鼠标左键向右下移动直至 E11 单元格松开左键，选中 A2:E11 区域。见图 2-10。

2. 将鼠标的光标定位于 A1:E11 区域中，按鼠标右键，在弹出的快捷菜单（见图 2-11）中，选中复制并单击。

3. 启动 SPSS。在 SPSS 数据编辑窗口，将鼠标光标移到左上角第一个单元格中，单击鼠标右键，在弹出的 SPSS 的快捷菜单（见图 2-12）中，选中 Paste（粘贴）单击，则在 SPSS 数据编辑窗口中，出现图 2-13 所示的数据文件。

需要指出的是，除非在 EXCEL 中输入的是字符或汉字，在 SPSS 中被默认为字符型名义测度变量，否则即使你对输入的数字定义成是文本型的，粘贴到 SPSS 数据编辑窗口中后，SPSS 将其默认为数值型尺度测度变量。

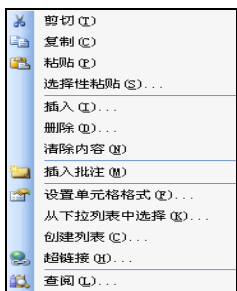


图 2-11 快捷菜单 1



图 2-12 快捷菜单 2

VAR00001	VAR00002	VAR00003	VAR00004	VAR00005
164.00	160.00	23.40	-1.40	172.00
164.00	162.00	22.20	-1.00	177.00
162.00	158.00	25.20	-1.80	180.00
160.00	159.00	22.00	4.00	167.00
158.00	167.00	24.20	1.00	160.00
158.00	161.00	21.50	0.00	162.00
156.00	160.00	22.60	-1.80	172.00
154.00	170.00	23.10	-4.00	180.00
151.00	152.00	21.20	2.00	159.00
148.00	155.00	20.40	2.00	158.00

图 2-13 粘贴后的变量、数据的形式

4. 按 Variable View 选项卡，在 Variable View 窗口中，修改变量名，定义变量类型和测度类型，有必要的话定义值标签和变量名标签，则可完成数据文件的转换。

虽然，它比直接在 SPSS 中录入数据作数据文件要快得多，但还不是两个不同数据分析软件之间最佳的数据文件转换的方法。

2.1.3.2 打开法

这是一种规范的做法，具体步骤如下：

1. 按 File→Open→Data 的顺序打开 Open Data 对话框，见图 2-14。
2. 利用操作系统中掌握的查找文件的方法，选择 Excel 数据文件所在的文件夹位置。
3. 在 File of type: 的下拉菜单中（见图 2-15）选择需要打开的文件类型为 Excel(*.xls,*.xlsx,*.xlsm)。

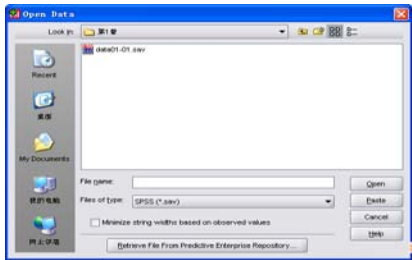


图 2-14 Open Data 对话框

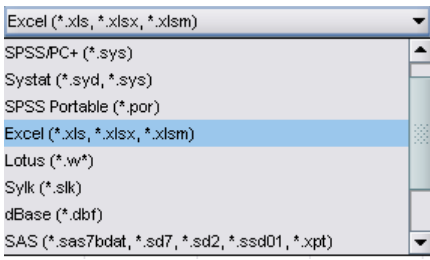


图 2-15 文件类型

4. 选中“第 2 章例 2.3 相关与回归数据资料.xls”，注意它应不在工作状态，如果工作簿正处在工作状态，应先关闭工作簿。

5. 单击 Open 按钮。出现 Opening Excel Data Source 对话框，见图 2-16。

6. 选择选中的 Excel 数据文件中第一行是否包含字段名。

由于在数据文件的第一行中有描述各列数据内容的字段名，因此 SPSS 将直接从第一行中获取变量名，故保持系统默认选项。否则单击该项选项，放弃这项选择。

7. 在 Worksheet 右侧下拉菜单中选择工作表。由于现在的数据正好放在第一张工作表中，因此也保持系统默认的选项。

如果调入的数据只是 Excel 数据文件中的一部分，则可以在 Range 右框中，用在 Excel 中定义区域的同样的方式，在此给出数据所在的区域。

在 Maximum width for string columns 右框中给出系统默认的最大变量数为 32767。

8. 单击 Continue 按钮，执行打开命令。在 SPSS 数据编辑窗口中出现的文件，见图 2-17。

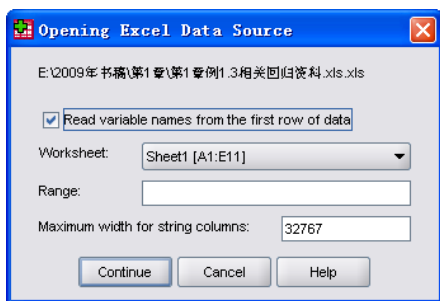


图 2-16 Opening Excel Data Source 对话框

	@13岁时身高	母亲身高	@13岁时足长	@13岁时骨龄	@17岁时身高
1	164	160	23.4	-1.4	172
2	164	162	22.2	-1.0	177
3	162	158	25.2	-1.8	180
4	160	159	22.0	4.0	167
5	159	167	24.2	1.0	160
6	158	161	21.5	0.0	162
7	156	160	22.6	-1.8	172
8	154	170	23.1	-4.0	180
9	151	152	21.2	2.0	159
10	149	155	20.4	2.0	158

图 2-17 出现在 SPSS 数据编辑窗口中的数据文件

注意，在 Excel 中，字段名的定义没有像 SPSS 中有首字符必须是字母或汉字的要求，所以凡不符合要求的 Excel 的字段名变成 SPSS 中相应的变量名时，在其最前面第一个字符的位置都被自动加了@。

正如前述，因为 Excel 数据文件中字段名下全部都是数字，因此，在 SPSS 中，变量被自动赋予数值型尺度测度。

9. 在 SPSS 中储存数据文件。

按 File→Save as 顺序打开 Save Sets as 对话框，见图 2-18。

按 Variables 按钮，打开 Save data as: Variables 对话框，见图 2-19。指定要储存在数据文件中的变量。本例全部选中。按 Continue 按钮，返回图 2-18。

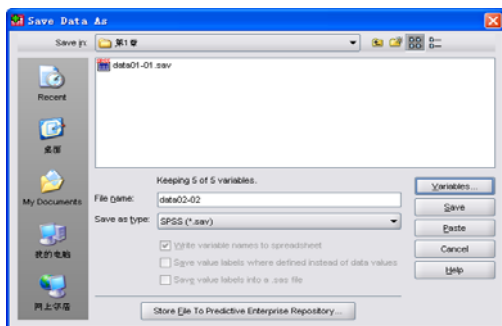


图 2-18 存储数据文件对话框

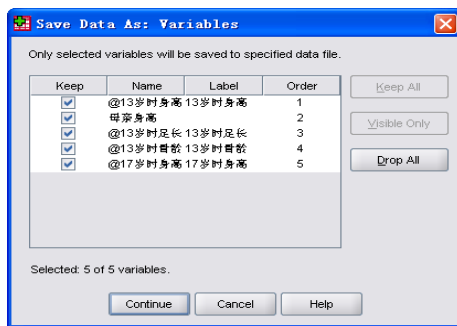


图 2-19 选择数据文件中包含的变量

在 File Name 右框中, 输入要储存的文件名为 data02-02。

Save as type (存储的类型) 采用系统默认的 SPSS (*.sav)。

单击 Save 按钮, 则在选定的目录中储存了名为 Data02-02.sav 的数据文件。

2.1.4 数据文件的合并

2.1.4.1 将另一个数据文件中的样品添加到现在的数据文件中

在规模稍大一些的科研中, 由于研究指标较多, 参与调研的被试对象人数也很多, 此时, 会收集到大量的研究分析用的数据资料, 数据整理工作将会占用很大一段时间, 为了尽快缩短这个过程, 需要较多的科研人员直接参与到数据录入工作中来, 所以在 SPSS 中录入数据资料时, 往往同时多人在做录入数据资料工作。这就需要在做最后的数据分析前, 要将多人的作业合并到一个主数据文件中。

例 2.4 根据统一规范的数据文件的变量设计后, 研究人员小张已将例 2.2 中前 100 位被试对象的调查问卷的结果录入 Data02-03.sav 中, 而小王将余下的 86 名被试对象的调查问卷的结果录入 Data02-04.sav 中, 现要将 Data02-04.sav 中的数据资料合并到 Data02-03.sav 中, 以便后续的统计计算分析工作。

由于这两个数据文件具有完全相同的变量结构, 即有相同的变量数量和变量定义, 因此可用以下两种方法实现数据文件的合并。

1. 复制法

从 SPSS14.0 起, 已可以同时打开多个数据编辑窗口, 所以当数据文件中具有完全相同的变量结构时, 复制法是一种最简捷的方法。

具体操作步骤如下:

(1) 依次单击 File→Open→Data, 打开 Open Data 对话框 (见图 2-14)。

(2) 利用操作系统中掌握的查找文件的方法, 找到 Data02-03.sav 数据文件, 选中并双击, 则打开的 Data02-03.sav 数据文件出现在 SPSS 新增的一个数据编辑窗口中。

(3) 同 (2) 一样, 找到并在 SPSS 新增的一个数据编辑窗口中打开 Data02-04.sav。则在显示器上出现如图 2-20 所示的画面。

(4) 在打开 Data02-04.sav 的数据编辑窗口中, 将鼠标光标移到数据显示区左上角的第一个单元格并单击, 选中该单元格。再将鼠标光标移到第 86 行最后一列交叉形成的单元格中, 按 Shift+单击, 则选中所有数据区域。将鼠标光标停在该区, 并单击右键, 在弹出的快捷菜单中, 见图 2-12, 选中 Copy 并单击。将选中的数据区域的内容存放在计算机内存的写字板中。

(5) 选中 Data02-03.sav 数据编辑窗口, 将鼠标光标移到数据显示区 101 行第一个单元格上, 单击鼠标右键, 在弹出的快捷菜单中选中 Paste 并单击, 则 Data02-04.sav 的所有数据资料, 都被粘贴在 Data02-03.sav 数据资料的下面, 完成两个数据文件的合并工作。

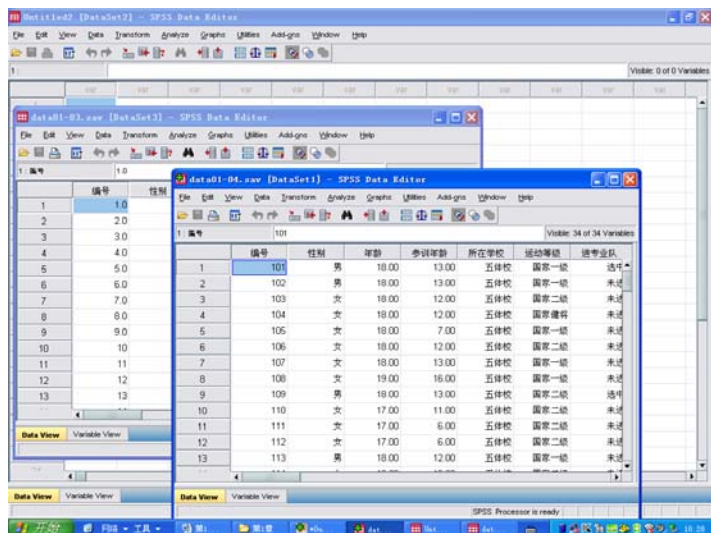


图 2-20 同时打开两个数据文件

(6) 按 File→Save as 顺序, 打开储存文件的对话框(见图 2-18), 在 File Name 后框中输入储存文件的地址和文件名。

(7) 单击 Save 按钮即可得到按(6)存取的磁盘文件。

2. 菜单法

其具体做法步骤如下:

(1) 在 SPSS 数据编辑窗口中打开将存放合并数据的主数据文件 Data02-03.sav。

(2) 按 Data→Merge Files→Add Cases 顺序, 打开 Add Cases 对话框, 见图 2-21。

(3) 选择合并数据文件的来源。

如果要合并的数据文件也已打开, 则选择第一个选择项。对于非 SPSS 数据文件要合并到 SPSS 的主数据文件中来, 则必须先在 SPSS 中打开。

由于本例中要被合并的数据文件 Data02-04.sav, 已是 SPSS 的数据文件, 所以合并时, 可以不用预先打开, 直接选用第二个选择项外部 SPSS 数据文件 (An external SPSS data file) 即可。在该选择项的下框中, 用 Browse(浏览)选定需要合并的文件名, 见图 2-21。单击 Continue 按钮, 弹出如图 2-22 所示的对话框。

(4) 选择变量名。

由于本例中, 两个数据文件的变量结构全部相同, 所以可以在本对话框中不作任何选择, 直接单击 OK 按钮即可。则在打开的 Data02-03.sav 中, 我们可以看到在 Data02-04.sav 中的 86 个样品被加到了数据文件的最后, 总样品数已变成 186 个。

(5) 按复制法中(6)、(7)步骤可得到合并后的磁盘数据文件。

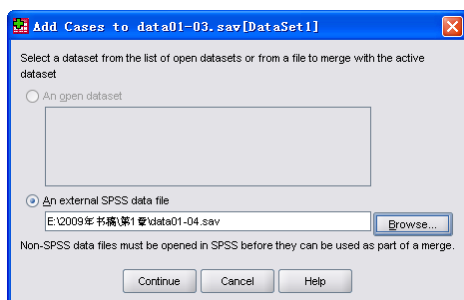


图 2-21 向 Date02-03 中增加样品对话框

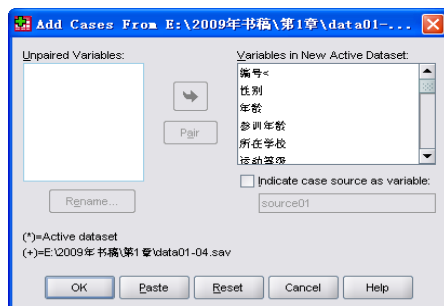


图 2-22 选择变量名

2.1.4.2 将另一个数据文件中的新变量添加到当前的数据文件中

例 2.5 在公路自行车多日赛中，运动员的最终成绩是由每段成绩累加而成。由于比赛开始后不能再有运动员报名参赛，所以根据最终技术会上确认的报名名单，可先建立一个包括号码、姓名、队名、UCI 代码在内的参赛运动员基本信息库，而在为每段赛后所建的成绩库中，只需有号码和每段成绩即可，这样只要通过参赛号码的连接，就可从基本信息库中取到打印输出中所需要的基本信息。现从 2007 年环海南国际公路自行车多日赛的成绩册中抽取部分运动员的基本信息及第一、第二赛段的成绩，分别存放在 data02-05.sav、data02-06.sav、data02-07.sav 三个 SPSS 数据文件中，目标是使用 SPSS 中提供的过程、程序将它们一同合并到一个同时含有这三个数据文件的数据文件 data02-08.sav 中。

实现这一目标的具体步骤如下：

1. 在 SPSS 中，按 File→Open→Data 顺序，打开 Open Data 对话框，见图 2-14。用上例中介绍的方法选中并打开 data02-05.sav 数据文件。

重复上面的做法，再依次打开 data02-06.sav、data02-07.sav 数据文件。见图 2-23、图 2-24、图 2-25。

2. 按号码排序。

有两种做法可实现对数据文件按号码排序：

- 按 Data→Sort Cases 顺序打开 Sort Cases 对话框，见图 2-26。将号码变量拖曳到 Sort by:框中，采用系统默认的 Ascending（升序）进行排序，单击 OK 按钮，实现对原数据文件中的样品进行按号码的（ASCII）顺序由小到大进行重排。

- 在打开数据文件的 SPSS 数据编辑窗口中，单击号码变量名，选中号码所在列，单击鼠标右键，在弹出的快捷菜单中，见图 2-27，选中 Sort Ascending 并单击，实现对原数据文件中的样品进行按号码的（ASCII）顺序由小到大进行重排。

3. 按 Data→Merge Files→Add Variables 顺序，打开 Add Variables 对话框，见图 2-28。

4. 选择合并数据文件的来源。

因为要合并的数据文件都已打开，所以选择第一个选择项 An open dataset，在其下框

中选择 data02-06.sav。按 Continue 按钮，弹出如图 2-29 所示的对话框。

	号码	姓名	队名
1	5	Bondariew Bogdan	INT
2	124	Bozhko Andrey	KAZ
3	23	Brooks Brendan	FRF
4	121	Chernyshov Liya	KAZ
5	6	Chmielewski	INT
6	24	Cridland Luke	FRF
7	32	Davel Shaun	AMO
8	101	Galimzyanov	RUS
9	21	Herize Peter	FRF
10	111	Hungerbuehler	SUI
11	26	Lyte Robert	FRF
12	171	Ma Haijun	CHN
13	92	Maartens Jeremy	RSA
14	42	Okazaki Kazuya	JPN
15	3	Osinski Marcin	INT
16	4	Radosz Robert	INT
17	115	Schelling Sven	SUI
18	172	Song Baoqing	CHN
19	64	White Bradley	MPC
20	174	Zou Rongxi	CHN

图 2-23 data02-05.sav

	号码	第一段成绩
1	101	0:03:45.08
2	171	0:03:45.72
3	26	0:03:46.65
4	42	0:03:46.88
5	5	0:03:47.82
6	64	0:03:48.43
7	21	0:03:48.57
8	111	0:03:49.32
9	32	0:03:49.56
10	115	0:03:50.06
11	92	0:03:50.36
12	23	0:03:50.56
13	172	0:03:50.77
14	3	0:03:51.07
15	24	0:03:51.72
16	174	0:03:52.17
17	6	0:03:54.10
18	4	0:03:54.27
19	124	0:03:54.58
20	121	0:03:55.55

图 2-24 data02-06.sav

	号码	第二段成绩
1	101	4:08:11.12
2	111	4:08:13.23
3	115	4:08:11.32
4	121	4:08:14.31
5	124	4:08:12.56
6	171	4:08:11.13
7	172	4:08:15.45
8	174	4:08:23.33
9	21	4:08:34.22
10	23	4:08:23.43
11	24	4:08:12.43
12	26	4:08:12.44
13	3	4:08:13.31
14	32	4:08:15.21
15	4	4:08:24.11
16	42	4:08:23.22
17	5	4:08:22.12
18	6	4:08:24.13
19	64	4:08:25.32
20	92	4:08:33.11

图 2-25 data02-07.sav

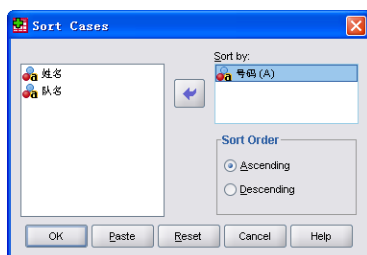


图 2-26 样品排序对话框

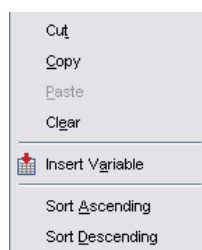


图 2-27 快捷菜单

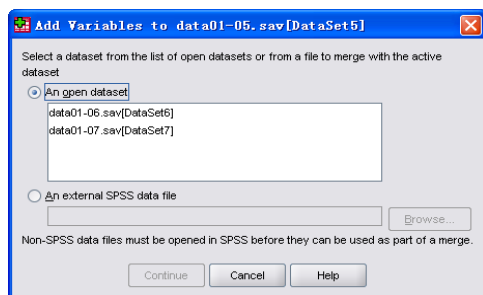


图 2-28 Add Variables 对话框

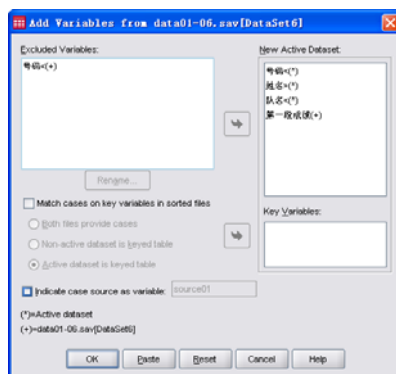


图 2-29 合并数据对话框

在 Exclude Variable（排除变量）下框中列出的是被合并数据文件中与当前数据文件

中同名的变量，它在合并文件中要从新变量中被排除。对话框中的*号代表在当前工作数据文件中的变量，+号代表要合并的数据文件中的变量。如果想在合并文件中包含一个与排除变量完全一样的名字，可以对它 **Rename**，并增加到包含的变量列表中。

在 **New Active Dataset**（合并数据文件中包含的新变量）下框中列出的是在新的合并工作数据文件中包含的所有不相同的变量名。

5. 选择 **Match cases on key variables in sorted files**（在排序文件的关键变量上的配对样品）选项。此复选项中有三个选项：

- **Both files provide cases**（由两个文件提供样品）选项，将选中的要合并的数据文件的样品添加到当前工作的数据文件中。

- **Non-active dataset is keyed table**（非工作数据文件是关键表）选项，只有与当前数据文件中关键变量相同的外部文件中的样品的观察值添加到当前工作文件中成为新变量。

- **Active dataset is keyed table**（工作数据文件是关键表）选项，将关键变量相同的工作数据文件中的样品的其他观察值作为新变量添加到外部数据文件中。

关于非工作数据文件是关键表的说明。关键表或表查找文件是一个文件，在这个文件中数据提供给每个“样品”，可被其他数据文件中的多重样品应用。例如，如果一个文件包含个人家庭成员的信息（如性别、年龄、教育），另一个文件包含总体家庭信息（如总收入、家庭人数、家庭地址），可用家庭数据文件作为一个表查找文件，及在合并数据文件中为每个家庭成员使用共同的家庭数据。

根据本列的具体目的，显然要选择 **Both files provide cases** 选项。

6. 选择关键变量。

本例需要选择关键变量，步骤如下：

在 **Exclude Variable**（排除变量）下框中选择 **号码**变量，单击 **Key Variables** 按钮：矩形框左侧的右移箭头按钮，将 **号码**变量定义为关键变量。

关于 **Key Variables**（关键变量）选项的说明：如果在数据文件中一些样品与另一个数据文件中没有匹配样品，即一些样品在另一个数据文件中是缺失的，使用关键变量定义并正确地从两个数据文件中匹配样品。此外，还能使用表里的关键变量查找文件。

- 在两个数据文件中关键变量必须有相同的名字。

- 两个数据文件必须按关键变量进行由小到大的升序排序，并且在关键变量列表中的变量的顺序必须同排序中顺序相同。

- 在关键变量上不匹配的样品包含在合并文件中，但同另一个文件中的样品不合并。不匹配样品只包含正在使用文件中的变量值，其他文件中的变量包含系统缺失值。

7. **Indicate case source as variable**（为每个样品指出来源的数据文件）选项，不是必选项。如果选择此项，其名称可在 **Indicate case source as variable** 的右框中定义，也可使

用系统默认给定的“source01”。则在合并的数据文件中，将增加一个新变量，其值 0 表示样品来源于当前数据文件，1 表示样品来源于外部数据文件。在本例中此项可以不选。

8. 单击 OK 按钮，则在当前工作的数据文件 data02-05 中，将 data02-06 中第一段成绩作为新变量添加进来。

重复上述步骤，同样可将 data02-07 中的第二段成绩作为新变量添加到当前工作的数据文件 data02-05 中来，见图 2-30。

	号码	姓名	队名	第一段成绩	第二段成绩
1	101	Galimzyanov	RUS	0:03:45.08	4:08:11.12
2	111	Hungerbuehler	SUI	0:03:49.32	4:08:13.23
3	115	Schelling Sven	SUI	0:03:50.06	4:08:11.32
4	121	Chernyshov Liya	KAZ	0:03:55.55	4:08:14.31
5	124	Bozhko Andrey	KAZ	0:03:54.58	4:08:12.56
6	171	Ma Haijun	CHN	0:03:45.72	4:08:11.13
7	172	Song Baoqing	CHN	0:03:50.77	4:08:15.45
8	174	Zou Rongxi	CHN	0:03:52.17	4:08:23.33
9	21	Herize Peter	FRF	0:03:48.57	4:09:12.12
10	23	Brooks Brendan	FRF	0:03:50.56	4:08:34.22
11	24	Cridland Luke	FRF	0:03:51.72	4:08:23.43
12	26	Lyte Robert	FRF	0:03:46.65	4:08:12.43
13	3	Osinski Marcin	INT	0:03:51.07	4:08:12.44
14	32	Davel Shaun	AMO	0:03:49.56	4:08:13.31
15	4	Radosz Robert	INT	0:03:54.27	4:08:15.21
16	42	Okazaki Kazuya	JPN	0:03:46.88	4:08:24.11
17	5	Bondariw Bogdan	INT	0:03:47.82	4:08:23.22
18	6	Chmielewski	INT	0:03:54.10	4:08:22.12
19	64	White Bradley	MPC	0:03:48.43	4:08:24.13
20	92	Maartens Jeremy	RSA	0:03:50.36	4:08:25.32

图 2-30 合并后的数据文件

9. 按 File→Save as 顺序，打开储存文件的对话框，见图 2-18，在 File Name 后框中输入 data02-08，则在当前目录下，存入一个名为 data02-08.sav 的 SPSS 合并数据文件。

2.1.5 数据文件的转置和重新构建

有时，在建立数据文件的过程中，根据分析的需要需转置变量。此时，不需要重新录入数据，只需用 SPSS 中提供的数据转置功能或重新构建功能就可以得到正确的数据文件。

2.1.5.1 数据文件的转置

例 2.6 某医院用中药治疗 9 例再生障碍性贫血患者，其血红蛋白（g/L）变化的数据见表 2-5。在 SPSS 中，建立的数据文件，见图 2-31，并存放在 data02-09.sav 中。

表 2-5 9 例再生障碍性贫血患者治疗前后的血红蛋白值

编号	1	2	3	4	5	6	7	8	9
治疗前	68	65	55	75	50	70	76	65	72
治疗后	128	82	80	112	125	110	85	80	105

	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	VAR00007	VAR00008	VAR00009
1	68.00	65.00	55.00	75.00	50.00	70.00	76.00	65.00	72.00
2	128.00	82.00	80.00	112.00	125.00	110.00	85.00	80.00	105.00

图 2-31 9 例再生障碍性贫血患者的血红蛋白值

如图 2-31 所示,这种数据的录入方式,没有考虑 SPSS 对变量数据的存放要求,所以,用它来进行统计分析会得到错误的结论。

对于类似本例的同一个被试对象在试验前后获得的数据资料,在 SPSS 中,一般要做相关的两个样本的差异性比较,因此,在 SPSS 中建立数据文件时,要按图 2-32 所示的方式建立配对试验设计时的数据文件,样例结果存放在 data02-10.sav 中。

因此,必须对其进行转置。具体操作步骤如下:

1. 打开数据 data 02-9.sav, 按 Data→Transpose 顺序, 展开 Transpose 对话框, 见图 2-33。
2. 在左侧变量名源框中, 选择全部变量, 按中间的向右箭头按钮, 将变量移入 Variable(s): 框中。
3. 单击 OK 按钮执行, 则在新的 SPSS 数据编辑窗口出现转置的数据文件, 见图 2-34。

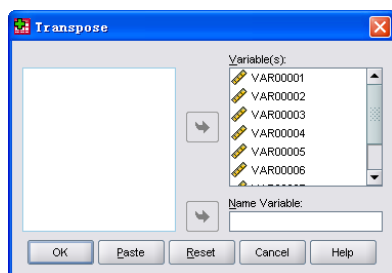


图 2-33 Transpose 对话框

	治疗前	治疗后
1	68.00	128.00
2	65.00	82.00
3	55.00	80.00
4	75.00	112.00
5	50.00	125.00
6	70.00	110.00
7	76.00	85.00
8	65.00	80.00
9	72.00	105.00

图 2-32 自身比较试验设计时的数据文件

	CASE_LBL	var001	var002
1	VAR00001	68.00	128.00
2	VAR00002	65.00	82.00
3	VAR00003	55.00	80.00
4	VAR00004	75.00	112.00
5	VAR00005	50.00	125.00
6	VAR00006	70.00	110.00
7	VAR00007	76.00	85.00
8	VAR00008	65.00	80.00
9	VAR00009	72.00	105.00

图 2-34 转置结果

对转置后的数据文件,做适当的编辑,如删除 CASE_LBL 变量,修改 var001 为治疗前,修改 var002 为治疗后,即可得到图 2-32 的所示的数据文件。

2.1.5.2 数据文件的重新构建

例 2.7 在研究鼻咽癌患者和非鼻咽癌患者在血清病毒 VCA-LOGA 抗体滴度方面有无差异, 实验人员用免疫酶法对 13 名鼻咽癌患者和 13 名非鼻咽癌患者的血清病毒 VCA-LOGA 抗体滴度分别进行了测定, 并将测试结果录入 SPSS 中, 建成了数据文件, 见图 2-35, 数据存放在 data02-11.sav 中。

	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	VAR00007	VAR00008	VAR00009	VAR00010	VAR00011	VAR00012	VAR00013
1	鼻咽癌患者	5.00	20.00	40.00	80.00	80.00	80.00	160.00	160.00	320.00	640.00	1280.00	
2	非鼻咽癌患者	5.00	10.00	10.00	20.00	20.00	20.00	40.00	40.00	80.00	80.00	80.00	160.00

图 2-35 鼻咽癌患者和非鼻咽癌患者血清病毒 VCA-LOGA 抗体滴度数据文件

上图显示的数据文件显然不符合 SPSS 的要求, 也无法进行后续的分析工作。

对于类似本例的数据资料, 在试验设计中把它称为试验对照或成组设计。在 SPSS 中, 一般要做独立的两个样本的差异性比较, 因此, 在 SPSS 中建立数据文件时, 要按图 2-36 所示的方式建立成组比较试验设计时的数据文件, 样例结果存放在 data02-12.sav 中。

因此, 必须对其进行重新构建。具体操作步骤如下:

1. 打开数据文件中 data02-11.sav, 按 Data→Restructure 顺序, 展开 Restructure 对话框, 见图 2-37。

	组别	抗体滴度
1	1	5.00
2	1	20.00
3	1	40.00
4	1	80.00
5	1	80.00
6	1	80.00
7	1	160.00
8	1	160.00
9	1	320.00
10	1	320.00
11	1	640.00
12	1	1280.00
13	1	1280.00
14	2	5.00
15	2	10.00
16	2	10.00
17	2	20.00
18	2	20.00
19	2	20.00
20	2	40.00

图 2-36 成组比较试验设计时的数据文件

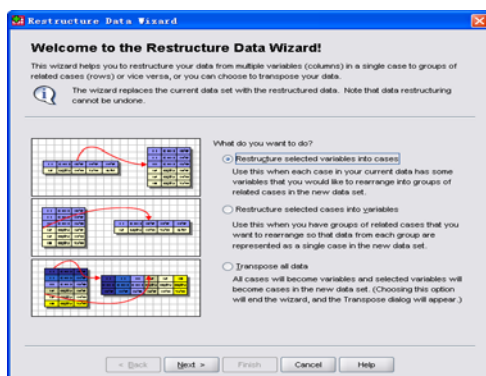


图 2-37 重建数据结构向导主对话框

在这个对话框中, 有三个选择项, 第一个选项是 Restructure selected variables into cases, 将选择的变量转换成样品(记录), 即列变行。第二个选项是 Restructure selected cases into variables, 将选择的样品变成变量, 即行变列。第二个选项是 Transpose all data, 对所有的数据进行转置。

本例的数据文件由于有不需要转换的数据, 所以选择第一个选项 Restructure selected variables into cases。

2. 单击 Next 按钮, 进入如图 2-38 所示的对话框。

如果重新构建的变量组只有一个, 选择第一个选项 One, 如果重新构建的变量组不

止一个，选择第二个选项 **More than one**。

在本例中，由于重新构建的变量组只有一个，所以选择第一个选项 **One**。

3. 单击 **Next** 按钮，进入如图 2-39 所示的对话框。

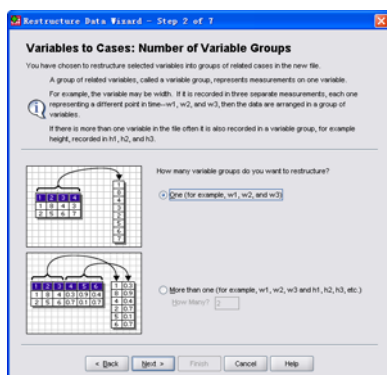


图 2-38 原始数据与第 2 步对话框

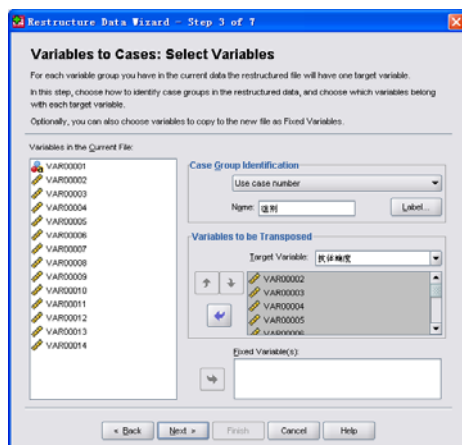


图 2-39 原始数据与第 3 步对话框

在 **Case Group Identification** 栏的下拉列表的选项中有三个选择项，用来确定在新的数据文件中变量的标志。

- 选择 **Use case number**，使用样品的顺序号，即在当前工作文件中的记录号来标志。
- 选择 **Use selected variable**，使用选择的变量来标志。
- 选择 **None**，不用标识变量。

本例，为了添置一个分组变量，因此，选择 **Use case number** 选项。并在 **Name** 框中，输入变量名为组别。

在 **Variable to be Transposed** 栏，用来确定需要转换的变量。

在 **Target variable:** 框中，定义变量名。本例定义的变量名为抗体滴度。

在 **Variables in the Current File** 下的变量名源框中，本例选择除第一个变量以外的其他所有变量，单击右箭头按钮，将它们移入 **Target variable:** 的下框中。

在 **Fixed Variable(s)** 下框，需要输入的是当前数据文件的变量名源框中不需要转换但还要出现在新数据文件中的变量。本例中，没有这个需求，故不输入任何变量。

4. 单击 **Next** 按钮，进入如图 2-40 所示的对话框。

在本对话框中，主要选择在新数据文件中生成多少个索引变量。

- 选择 **One**，生成一个索引变量。
- 选择 **More than one**，生成一个以上的索引变量。
- 选择 **None**，不生成索引变量。

本例中选择 **None**，不生成索引变量。

5. 单击 Next 按钮, 进入如图 2-41 所示的对话框。

Handling of Variables not selected 栏用来确定原始数据文件中未被选择的变量的处理方式。它有两个选项:

- 选择 Drop variable(s) from the new data file, 在新数据文件中不包括那些未被选择的变量。

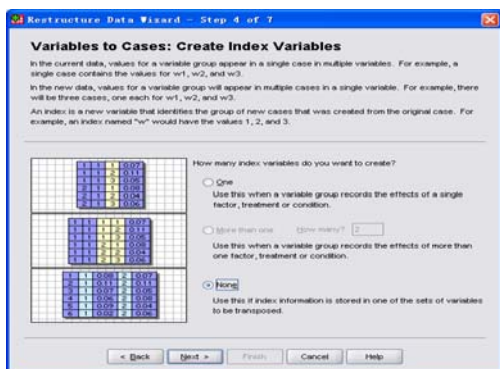


图 2-40 原始数据与第 4 步对话框

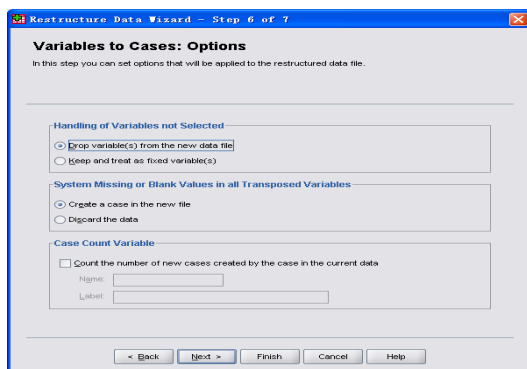


图 2-41 原始数据与第 6 步对话框

- 选择 Keep and treat as fixed variable(s), 将那些变量保留在新数据文件中并作为固定变量处理。

System Missing and Blank Values in all Transposed Variables 栏用来确定无效值的处理方式。它有两个选项:

- 选择 Create a case in the new file, 在新数据文件中生成一个样品。
- 选择 Discard the data, 删除这个数据。

Case Count Variable 栏用来确定是否在转换后的新数据文件中生成样品计数变量。它只有一个选项 Count the number of new case created by the case in the current data file, 系统默认为不选择。如果选择本项, 则在新的数据文件中生成计数变量。它记录由当前数据文件中的样品所建立的新样品的数量。

本例对本对话框中的内容可以不做任何选择。

6. 单击 Next 按钮, 进入如图 2-42 所示的最后一歩的对话框。

它决定你想做什么决定。包括两个选项:

- 选择 Restructure the data now, 则在当前数据文件中, 重新构建由上述设定条件产生的新文件。
- 选择 Paste the syntax generated by the wizard into a syntax window, 将由 Wizard 产生的语句粘贴

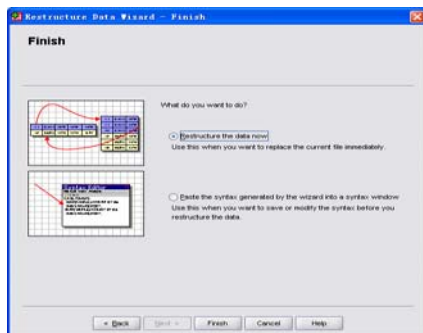


图 2-42 原始数据与第 7 步对话框

到语句窗口。当你没有准备好改变当前文件，当你想修改语句或保留它以便以后再用时，选择此项。

本例选择 **Restructure the data now** 选项，则当前数据文件被重新构建成新的数据文件，所建立的成组比较试验设计数据文件如图 2-36 所示。

关于重新构建数据文件中更详细的内容，可参阅由卢纹岱主编的《SPSS 统计分析（第4版）》第2章 2.3.4 节中的相关内容。

2.1.6 其他特殊数据文件的建立

在以上内容中，通过实例的形式已经介绍了几种常用数据文件的格式。此外，还会遇到以下一些特定条件下的数据文件的建立。

2.1.6.1 多维列表资料的数据文件的建立

在对四格表及四格表以上的多维列表的数据资料整理成 SPSS 的数据文件时，主要考虑的是在 SPSS 中至少要建立多少个变量，才能把数据资料的信息完整地反映出来，并能在 SPSS 中进行有效的分析。

变量个数取决于对观测结果影响的因素的多少和观测指标的多少。一般而言，变量数等于影响因素数加观测指标数。

变量的种类可以分为分类变量（含分组变量，名义测度）和观测指标变量（尺度测度）两种。

例 2.8 为了解男性和女性对淡啤酒、普通啤酒和黑啤酒的偏好有无差异，分别调查了男性饮啤酒者 1353 人和女性饮啤酒者 636 人，结果见表 2-6。试把它整理成 SPSS 的数据文件。

表 2-6 男性和女性饮啤酒者对三种类型的啤酒偏好的观察频数

	啤酒偏好			合计
	淡啤酒	普通啤酒	黑啤酒	
男性	352	284	717	1353
女性	293	133	210	636
合计	645	417	927	1989

在本例中，尽管表 2-6 中的数据较多，似乎头绪很多，无从下手，但仔细观察不难发现，表格边缘的合计数是由表中数据累加而得，因此，它们可以不用考虑用单独建立变量的方式来加以描述，重点应放在表格中间的数据上，观察不难发现，观测指标只有一个，就是被试者对啤酒类型的喜好，而影响表格中间饮啤酒者选择结果的因素有两个，一个是被试者的性别因素，另一个是被试者对啤酒的偏好因素。

因此，本例只需建立三个变量：两个分类（分组）变量和一个观测频数变量即可。

在 SPSS 中建立的数据文件，见图 2-43 和图 2-44，样例存放在 data02-13.sav 中。

	性别	啤酒类型	观测频数
1	1.00	1.00	352.00
2	1.00	2.00	284.00
3	1.00	3.00	717.00
4	0.00	1.00	293.00
5	0.00	2.00	133.00
6	0.00	3.00	210.00

图 2-43 变量及变量值

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	性别	Numeric	8	2		(0.00, 文)...	None	8	靠右	Nominal
2	啤酒类型	Numeric	8	2		(1.00, 淡啤...	None	8	靠右	Nominal
3	观测频数	Numeric	8	2		None	None	8	靠右	Scale

图 2-44 变量定义

值得一提的是，在用这样的数据文件进行后续的统计分析时，先用观测频数变量作为权重变量，对数据文件进行加权处理，具体操作步骤，详见第 2 章 2.3.1.4 中的例 2.21。

2.1.6.2 单因素试验设计资料和多因素试验设计资料的数据文件的建立

在将单因素和多因素试验设计的试验结果整理成 SPSS 的数据文件时，与 2.6.1 中一样，也要考虑至少要建立多少个才能描述试验的数据资料。

	种子	施肥	区组	玉米产量
1	1.00	1.00	1.00	17.00
2	1.00	2.00	1.00	11.00
3	1.00	3.00	1.00	12.00
4	2.00	1.00	1.00	19.00
5	2.00	2.00	1.00	20.00
6	2.00	3.00	1.00	17.00
7	3.00	1.00	1.00	19.00
8	3.00	2.00	1.00	10.00
9	3.00	3.00	1.00	9.00
10	1.00	1.00	2.00	15.00
11	1.00	2.00	2.00	14.00
12	1.00	3.00	2.00	8.00
13	2.00	1.00	2.00	13.00
14	2.00	2.00	2.00	19.00
15	2.00	3.00	2.00	16.00
16	3.00	1.00	2.00	18.00
17	3.00	2.00	2.00	8.00
18	3.00	3.00	2.00	8.00
19	1.00	1.00	3.00	13.00

图 2-45 变量及变量值

单因素试验设计与成组设计一样，只需建立两个变量，一个是分组变量，用来存放因素所分的不同水平，另一个是试验结果变量，用来存放试验中测试指标的值，参见例 2.7。

在多因素试验设计中，变量的个数同影响试验结果的因素个数和试验中所要测试的指标个数有关。通常变量的个数应为影响试验结果的因素个数与所要测试的指标个数的和。在有重复试验的情况下，变量的个数中还应包括一个重复测试变量。

例 2.9 为研究玉米品种 (A) 与施肥因素 (B) 对玉米产量的影响，将玉米种子分成三种，即三个水平 A1、A2、A3，施肥因素也分三个水平，即 B1、B2、B3。将土壤质地相同的田间分成 36 块。每个试验方案重复 4 次试验，结果见表 2-7。

表 2-7 玉米种子与施肥因素对玉米产量影响的试验结果

处理		区组			
		1	2	3	4
A1	B1	17.0	15.0	13.0	14.0
	B2	11.0	14.0	13.0	12.0
	B3	12.0	8.0	8.0	9.0
A2	B1	19.0	13.0	11.0	11.0
	B2	20.0	19.0	13.0	14.0

(续表)

处理		区组			
		1	2	3	4
	B3	17.0	16.0	18.0	16.0
A3	B1	19.0	18.0	16.0	17.0
	B2	10.0	8.0	10.0	8.0
	B3	9.0	8.0	7.0	7.0

在本例中，试验指标为玉米产量，影响玉米产量的因素有玉米品种（A）与施肥因素（B），它又是重复试验，所以，在 SPSS 中建立数据文件时，需要设立 4 个变量：一个试验指标变量，两个因素变量和一个区组重复试验变量。

在 SPSS 中建立的数据文件，见图 2-45 和图 2-46，样例存放在 data02-14.sav 中。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	种子	Numeric	8	2		{1.00, A1}...	None	8	Right	Nominal
2	施肥	Numeric	8	2		{1.00, B1}...	None	8	Right	Nominal
3	区组	Numeric	8	2		{1.00, 区组1}...	None	8	Right	Nominal
4	玉米产量	Numeric	8	2		None	None	8	Right	Scale

图 2-46 数据文件中各变量的定义

2.1.6.3 多元正态总体均值等于常数向量检验时的数据文件的建立

在建立 SPSS 的数据文件中，有关多元正态总体均值等于常数向量检验时建立的数据文件是个特例，之所以说它是特例，是因为在第 1 章中涉及变量的定义时，已经谈到变量是可以取各种各样值的量，它是可变的，如果不变，都取一样的值，那它就不是变量，而是常量。一般而言，常量在极大多数的统计分析中是没有多大用处的，但对本类型的问题来说，它绝对是个例外。

因此，对本类型的问题，在 SPSS 中建立数据文件时，需要建立的变量个数为指标个数加一个存放常量的“变量”。

例 2.10 在研究排汗量与体内钠含量与钾含量关系的研究中，测试了 20 名健康女性在这三个指标上的具体数据，见表 2-8。使用 SPSS 中的程序分析这三个指标的均值是否为 4、50、10，请建立恰当的数据文件。

表 2-8 20 名女子排汗量与体内钠含量与钾含量记录表

编号	排汗量	钠含量	钾含量	编号	排汗量	钠含量	钾含量
1	3.7	48.5	9.3	11	3.9	36.9	12.7
2	5.7	65.1	8	12	4.5	58.8	12.3
3	3.8	47.2	10.9	13	3.5	27.8	9.8
4	3.2	53.2	12	14	4.5	40.2	8.4

(续表)

编号	排汗量	钠含量	钾含量	编号	排汗量	钠含量	钾含量
5	3.1	55.5	9.7	15	1.5	13.5	10.1
6	4.6	36.1	7.9	16	8.5	56.4	7.1
7	2.4	24.8	14	17	4.5	71.6	8.2
8	7.2	33.1	7.6	18	6.5	52.8	10.9
9	6.7	47.4	8.5	19	4.1	44.1	11.2
10	5.4	54.1	11.3	20	5.5	40.9	9.4

在本例中, 由于在同一个被试对象身上测试的指标为 3 个, 因此必须设立三个变量来存放这些测试指标的值。另外, 为能使用一般线性模型中的多元方差分析的方法来做多元正态总体均值等于常数向量的检验, 必须另加一个相当于分组变量但其值都为同一个数的常量“变量”。此例中用“1”来表示(见图 2-47)。

在 SPSS 中建立的数据文件, 见图 2-47, 样例存放在 data02-15.sav 中。

	类别	排汗量	钠含量	钾含量
1	1	3.70	48.50	9.30
2	1	5.70	65.10	8.00
3	1	3.80	47.20	10.90
4	1	3.20	53.20	12.00
5	1	3.10	55.50	9.70
6	1	4.60	36.10	7.90
7	1	2.40	24.80	14.00
8	1	7.20	33.10	7.60
9	1	6.70	47.40	8.50
10	1	5.40	54.10	11.30
11	1	3.90	36.90	12.70
12	1	4.50	58.80	12.30
13	1	3.50	27.80	9.80
14	1	4.50	40.20	8.40
15	1	1.50	13.50	10.10
16	1	8.50	56.40	7.10
17	1	4.50	71.60	8.20
18	1	6.50	52.80	10.90
19	1	4.10	44.10	11.20
20	1	5.50	40.90	9.40

图 2-47 变量及变量值

2.1.6.4 计算变异系数时数据文件的建立

在生物学、生理学等学科的研究中, 变异系数是一个比较常用的描述数据资料离散程度的指标, 在 SPSS 中, 计算变异系数的程序较少, 参见第 1 章 1.4.1.4 中的表 1-10。其实例计算参见第 2 章例 2.43。由于 Descriptive Statistics 中的 Ratio 程序并不是专门为计算变异系数设计的, 因此, 用它来计算变异系数时, 对数据文件中变量的设立也有特殊的要求。除要根据计算变异系数的指标个数建立相应数量的变量外, 另外还要增设一个当作分母使用的数字型常量 1 的“变量”。

例 2.11 实测 10 名 13 岁男孩的身高与足长的资料, 见表 2-9, 为比较身高与足长的

离中趋势。试建立 SPSS 的数据文件。

在本例中，由于需要计算变异系数的变量有两个：身高和足长，并在一次计算中同时完成，因此，在数据文件中共需建立三个变量：变量类别、测试值和系数。

在变量类别中，定义其值（Values）为：1=身高，2=足长，测度水平为名义。在测试值变量中，存放身高、足长的测试值，测度水平为尺度。在系数中，存放一组常量 1。

在 SPSS 中建立的数据文件，见图 2-48，样例存放在 data02-16.sav 中。

2.1.6.5 对含有多项名义数据的资料进行二项逻辑斯蒂回归时数据文件的建立

在一些特定的场合，有时名义变量的取值不是两个，而是多个，例如，描述婚姻状况的变量，一般分为四种类型：未婚、有配偶、丧偶和离婚；又如，籍贯，可以是北京、上海、江苏、辽宁等省、直辖市等中的一个地方，它是一个典型的多项名义数据。

通常情况下，建立一个名义变量就可将名义数据的多个取值（多种状况），全部进行描述，例如，用 1、2、3 和 4 分别表示未婚、有配偶、丧偶和离婚，即可。

表 2-9 13 岁男孩身高与足长资料

	身高	足长
1	164.00	23.40
2	164.00	22.20
3	162.00	25.20
4	160.00	22.00
5	158.00	24.20
6	158.00	21.50
7	156.00	22.60
8	154.00	23.10
9	151.00	21.20
10	148.00	20.40
N	10	10

	变量类别	测试值	系数
1	1.00	164.00	1.00
2	1.00	164.00	1.00
3	1.00	162.00	1.00
4	1.00	160.00	1.00
5	1.00	158.00	1.00
6	1.00	158.00	1.00
7	1.00	156.00	1.00
8	1.00	154.00	1.00
9	1.00	151.00	1.00
10	1.00	148.00	1.00
11	2.00	23.40	1.00
12	2.00	22.20	1.00
13	2.00	25.20	1.00
14	2.00	22.00	1.00
15	2.00	24.20	1.00
16	2.00	21.50	1.00
17	2.00	22.60	1.00
18	2.00	23.10	1.00
19	2.00	21.20	1.00
20	2.00	20.40	1.00

图 2-48 数据结构

但如果要用二项逻辑斯蒂回归进行分析时，建立的数据文件中对名义变量的设立，同一般情况下建立数据文件时是有区别的，要参照以下的规定进行：

如果某个名义变量有多项值可取，假设有 k 项，则应将这名义变量拆分成 $k-1$ 个变量， M_1, M_2, \dots, M_{k-1} ，使得对每个 $i=1, \dots, k-1$ ， $(M_i=1, M_j=0, j=1, \dots, i-1, i+1, k-1)$ 代表第 i 个状况；并且以 $(M_j=0, j=1, \dots, k-1)$ 或 $(M_j=-1, j=1, \dots, k-1)$ 代表第 k 个状况。

例 2.12 1990 年第四次人口普查把婚姻状况分为四种类型：未婚、有配偶、丧偶和离婚。在过去的一年里，上海市 25 岁及以上的人中，各类婚姻状况的居民和死亡人数见表 2-10。在需要对这批数据做不同的年龄和婚姻状态对死亡率的影响的二项逻辑斯蒂回归之前，如何建立相应的数据文件。

表 2-10 各年龄组和各类婚姻状况的居民人数及死亡人数统计表

年龄	健在否	未婚	有配偶	丧偶	离婚
25~29	健在	318869	973796	944	6165
	死亡	349.1	417.1	4.1	11.4
30~34	健在	135589	1495515	2738	16333
	死亡	329.2	877.8	11.6	25.5
35~39	健在	52427	1354019	5383	17901
	死亡	213.9	1268.8	16.0	26.0
40~44	健在	21781	926994	9703	12343
	死亡	127.5	1299.2	31.5	27.9
45~49	健在	9242	603685	11666	7870
	死亡	68.7	1357.1	45.4	26.3
50~54	健在	6441	562182	24099	7191
	死亡	86.4	2107.4	130.1	38.3
55~59	健在	5847	680892	57015	9106
	死亡	99.4	4255.2	446.1	77.4
60~64	健在	4343	536601	89970	8768
	死亡	92.9	5868.7	1082.7	117.6
65~69	健在	3423	377263	123005	7615
	死亡	119.5	7240.7	2351.4	159.6
70 岁以上	健在	5292	354609	372391	7820
	死亡	434.8	20271.0	29842.0	414.3

由于应变量为死亡率，自变量（二项逻辑斯蒂回归中把它们称做协变量）为年龄和婚姻状况，在自变量中，婚姻状况是多项名义变量，因此，在建立数据文件的变量时，建立一个**健在否**的因变量，建立一个**年龄**变量，各组年龄可取组中值，即（年龄组下限+组上限）/2，最后一组年龄，取 70。而名义数据婚姻状况，由于它分 4 种状况，所以要建立 4-1 个名义变量，即 m1、m2 和 m3，四种不同的婚姻状况下的 m1、m2 和 m3 的取值分别如下：

(m1=1, m2=0, m3=0) 表示未婚;

(m1=0, m2=1, m3=0) 表示有配偶;

(m1=0, m2=0, m3=1) 表示丧偶;

(m1=0, m2=0, m3=0) 表示离婚。

当然, 用 (m1=-1, m2=-1, m3=-1) 定义离婚, 也是可以的。本例用 (m1=0, m2=0, m3=0) 表示离婚。

因此, 建立的数据文件的结构可参见图 2-49, 样例存放在 data02-17.sav 中。

	年龄段	年龄	健在否	m1	m2	m3	观察人数
1	1.00	27.00	1.00	1.00	0.00	0.00	318869.00
2	1.00	27.00	0.00	1.00	0.00	0.00	349.10
3	2.00	32.00	1.00	1.00	0.00	0.00	135589.00
4	2.00	32.00	0.00	1.00	0.00	0.00	329.20
5	3.00	37.00	1.00	1.00	0.00	0.00	52427.00
6	3.00	37.00	0.00	1.00	0.00	0.00	213.90
7	4.00	42.00	1.00	1.00	0.00	0.00	21781.00
8	4.00	42.00	0.00	1.00	0.00	0.00	127.50
9	5.00	47.00	1.00	1.00	0.00	0.00	9242.00
10	5.00	47.00	0.00	1.00	0.00	0.00	68.70
11	6.00	52.00	1.00	1.00	0.00	0.00	6441.00
12	6.00	52.00	0.00	1.00	0.00	0.00	86.40
13	7.00	57.00	1.00	1.00	0.00	0.00	5847.00
14	7.00	57.00	0.00	1.00	0.00	0.00	99.40
15	8.00	62.00	1.00	1.00	0.00	0.00	4343.00
16	8.00	62.00	0.00	1.00	0.00	0.00	92.90
17	9.00	67.00	1.00	1.00	0.00	0.00	3423.00
18	9.00	67.00	0.00	1.00	0.00	0.00	119.50
19	10.00	70.00	1.00	1.00	0.00	0.00	5292.00
20	10.00	70.00	0.00	1.00	0.00	0.00	434.80
21	1.00	27.00	1.00	0.00	1.00	0.00	973796.00
22	1.00	27.00	0.00	0.00	1.00	0.00	417.10
23	2.00	32.00	1.00	0.00	1.00	0.00	1495515.00

图 2-49 建立二项逻辑斯蒂回归的多项名义数据文件

2.2 数据资料的整理：频数分布表的制作

频数分布是根据资料内观察值数目的多少, 把整个资料分成若干个组, 并把每个观察值分别归到相应的组内, 统计各组频数得到的分布。

2.2.1 定性数据资料频数分布表的制作

定性数据资料可以按变量的属性或属性的不同程度进行分类, 然后根据各个体的具体表现, 分别归入相应的组内, 统计各组次数。

例 2.13 对例 2.2 进行实际调查后获得的数据资料见 data02-01.sav, 试分析接受调查者中, 运动等级的分布情况。

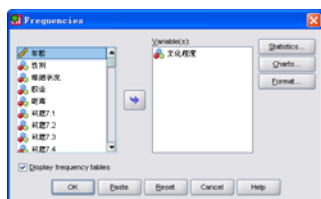


图 2-50 频数对话框

在 SPSS 中,分类变量的频数分布表的制作步骤如下:

1. 按 Analyze→Descriptive Statistics→Frequencies 顺序,展开 Frequencies 对话框,见图 2-50。在变量名表中,选择所要分析的运动等级变量,并将其移到 Variables: 矩形框中,其他保持系统默认选项。

2. 单击 OK 按钮运行,则在输出窗口中得到频数分布表,见表 2-11。

表 2-11 运动等级频数分布表

运动等级					
	Frequency	Percent	Valid Percent	Cumulative Percent	
Valid 国家健将	9	4.8	4.8	4.8	
国家一级	51	27.4	27.4	32.3	
国家二级	122	65.6	65.6	97.8	
国家三级	4	2.2	2.2	100.0	
Total	186	100.0	100.0		

从表 2-11 的频数分布表中,从左向右,依次列出的是变量值(国家健将、国家一级、国家二级、国家三级),各变量值对应的观察频数(9、51、122、4)、频率(4.8%、27.4%、65.6%、2.2%)、有效观察次数的百分比(同频率)、累计有效观察频数的百分比(4.8%、32.3%、97.8%、100%)。

最后一行列出的总和分别是总频数 186, 即有 186 位受访者接受了调查,列出的四个运动等级中,受访者 100% 进行应答,有效百分比也为 100%。

2.2.2 定量数据资料频数分布表的制作

为了对一批计量资料做更加深入的研究,有时需要将计量资料变成计数资料中的有序等级资料,这需要用频数分布表。

从例 2.13 中,不难发现,用 Frequencies 命令可求得一组观察值中,各个观察值出现的次数。它适合于分类变量。

但对计量资料而言,它并不希望统计一组测量值中,各个测试值出现了多少次,更多要看的是在一个个具体给定的区间中,观察值出现了多少次,如表 2-12 所显示的。

表 2-12 大气中 CO₂ 分布表

组限	天数
25~	39
50~	67
75~	64
100~	63
125~	45
150~	30
175~	17
200~	9
225~	7
250~	6
275~	5
300~	3
325~	6

因此，直接对原始数据资料调用 **Frequencies** 命令是达不到理想的效果的。

现结合实例来说明计量资料频数分布表的制作过程。

例 2.14 实测 100 名健康成年女子血清总蛋白含量，见表 2-13，原始数据做成的 SPSS 数据文件存放在 data02-18.sav 中。

表 2-13 100 名健康成年女子血清蛋白含量记录表（单位：g%）

7.43	7.88	6.88	7.80	7.04	8.05	8.05	6.97	7.12	7.35
7.95	7.56	7.50	7.88	7.20	7.20	7.20	7.43	7.12	7.20
7.50	7.35	7.88	7.43	7.58	6.50	7.43	7.12	6.97	6.80
7.35	7.50	7.20	6.43	7.58	8.03	6.97	7.43	7.35	7.35
7.58	7.58	6.88	7.65	7.04	7.12	8.12	7.50	7.04	6.80
7.04	7.20	7.65	7.43	7.65	7.76	6.73	7.20	7.50	7.43
7.35	7.95	7.35	7.47	6.50	7.65	8.16	7.54	7.27	7.27
6.72	7.65	7.27	7.04	7.72	6.88	6.63	6.73	6.73	7.27
7.58	7.35	7.50	7.27	7.35	7.35	7.27	8.16	7.03	7.43
7.35	7.95	7.04	7.15	7.27	7.72	8.43	7.50	7.65	7.04

现要对血清总蛋白含量数据资料制作频数分布表。

具体的制作步骤如下：

1. 确定组数 k

一般而言，数据资料的个数越多（即样本含量越大），所分组数也相应越多。但究竟分多少合适，还没有一个统一的定论。这里可以参照前苏联统计学家 H·A 马萨利金博士依其经验提出的根据样本含量来确定组数的方法，见表 2-14。

表 2-14 (苏)马萨利金博士的分组方法

样本含量 n	可分组数 k
30—60	5—8
60—100	7—10
100—200	9—12
200—500	11—16

本例可以分 10 组左右。

2. 确定组距 i

在制作频数分布表时，要遵循一个原则，任一个原始数据只能出现在该表的一个组中，不能同时兼属两个组中。这一原则指出，制作频数分布表时，组与组之间应该有明确的界限，即组限，每组的起点称组下限，而每组的止点称组上限。由此可见，对计量

资料而言, 组限应是闭、开区间, 而对计数资料而言, 组限应是闭区间。

组与组之间的间隔称为组距。各组间距离都相等的分组方法叫等组距法, 反之称不等组距法。这两种分组方法在实际工作中都经常用到, 由于原理是一样的, 所以, 本例使用等组距法。组距 i 由下式确定

$$i = \frac{R}{k}$$

式中 R 是两极差, 即原始数据中的最大值与最小值之差。组距一般取同上式计算结果相近的一个较完整的值。

因此, 本例的组距 $i = \frac{R}{k} = \frac{2}{10} = 0.2$ 。

3. 确定组限

制作频数分布表应遵循的第二个原则是所有的原始数据都应出现在所制作的频数分布表中。所以, 如果分组的组限按由小到大的顺序排列, 则第一组的下限应小于等于原始数据资料的最小值, 最后一组的上限应大于等于原始数据资料的最大值。反之则相反。

因此, 要制作本例的频数分布表, 首先要计算血清总蛋白含量资料的最大值与最小值。具体步骤如下:

(1) 在 SPSS 数据编辑窗口打开 data02-18.sav。

(2) 按 Analyze→Reports→Case Summaries 顺序打开 Case Summaries 的对话框, 见图 2-51。在左侧的源变量框中将变量血清总蛋白用单击鼠标左键的方式选中, 单击向右箭头按钮, 将其送入 Variables 框中。关闭 Display Cases 选择项, 单击 Statistics 按钮, 展开 Statistics 选择项对话框, 见图 2-52。将统计量 Number of Cases (样本含量)、Maximum (最大值)、Minimum (最小值)、Range (两极差) 移入右框中。单击 Continue 按钮, 返回图 2-51。单击 OK 按钮运行, 在输出窗中得到表 2-15 的输出结果。

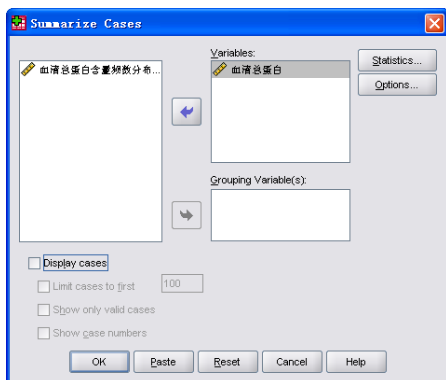


图 2-51 Case Summaries 对话框

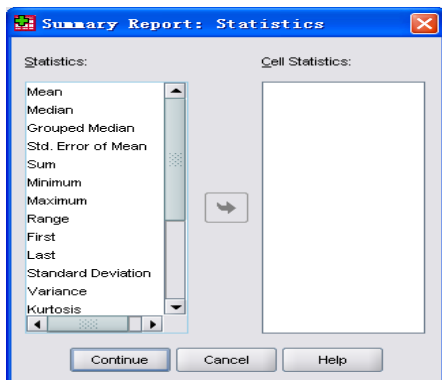


图 2-52 Statistics 对话框

由表 2-15 可见, 本例数据资料中样本含量为 100, 最大值为 8.43, 最小值为 6.43, 两极差 R 为 2.00。

在本例中, 组限按由小到大顺序排列, 故第一组的下限取 6.40, 它包含了最小值 6.43, 则第一组的近似上限为 6.60。

由此可得其后各组的下限依次为: 6.60, 6.80, 7.00, 7.20, 7.40, 7.60, 7.80, 8.00, 8.20, 由于分到预定的十组时, 最后一组的近似上限 8.40 小于最大值 8.43, 所以再顺延一组下限为 8.40。

4. 将原始数据换算成相应各组的组下限

在 SPSS 中有两种不同的做法。

(1) 利用 Transform 菜单下的 Compute Variable 过程来实现, 具体步骤如下:

① 按 Transform→Compute Variable 顺序打开 Compute Variable 对话框, 见图 2-53。在 Target Variable 的下框中输入目标变量名“组下限”。

② 在 Numeric Expression 的下框中输入第一组的下限值 6.40。

③ 单击 If 按钮, 打开 If Case 对话框, 见图 2-54。选择有条件限制的第二选择项 (Include if case satisfies condition), 给出使组下限=6.40 的条件为血清总蛋白<6.60, 单击 Continue 按钮返回上一窗口。

表 2-15 计算结果

Case Summaries

血清总蛋白

N	Minimum	Maximum	Range
100	6.43	8.43	2.00

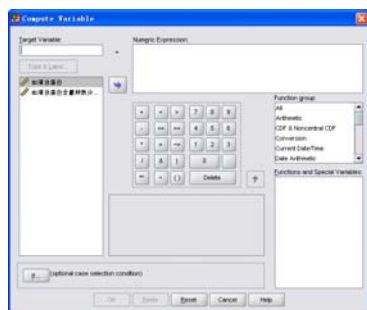


图 2-53 Compute 对话框

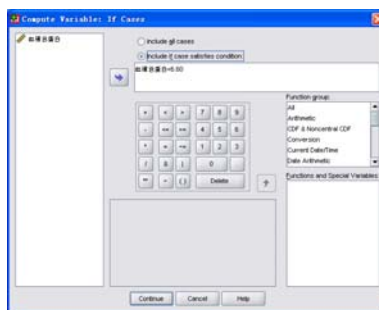


图 2-54 If Case 对话框

④ 单击 OK 按钮运行, 则凡血清总蛋白<6.40 的, 在其对应的组下限下被赋值为 6.40。显然, 组下限=6.60 的条件为血清总蛋白>=6.60 & 血清总蛋白<6.80; 组下限=6.80 的条件为血清总蛋白>=6.80 & 血清总蛋白<7.00; 组下限=7.00 的条件为血清总蛋白>=7.00 & 血清总蛋白<7.20; 组下限=8.20 的条件为血清总蛋白>=8.20 & 血清总蛋白<8.40; 组下限=8.40 的条件为血清总蛋白>=8.40 & 血清总蛋白<8.60。

重复上述过程，直到最后一组计算结束。则在组下限中得到血清总蛋白对应的转换值。

(2) 利用 Transform 菜单下的 Into Different Variables 过程来实现。具体步骤如下：

① 按 Transform→Recode into Different Variables 顺序，打开 Recode into Different Variables 对话框，见图 2-55。

② 将血清总蛋白移入 Numeric Variable→Output Variable 框下，在 Output Variable 的 Name: 框下输入“组下限”，在 Label: 框下输入“血清总蛋白含量频数分布表”，单击 Change 按钮，完成输入变量与输出变量名的设置。

③ 单击 Old and New Values 按钮，打开 Old and New Values 对话框，见图 2-56。选择 Range，在 through 左侧的框中输入 6.40，在 through 右侧的框中输入 6.599，在 New Value 的选项中，选择 Value 并在其后的框中，输入 6.40，然后单击 Add 按钮，将其添加到 Old→New 框中。再在 through 的左、右框中输入 6.60 和 6.799，在 Value 后的框中输入 6.60，单击 Add 按钮，将其添加到 Old→New 框中。每次增加 0.2 的组距，重复上述过程直至将最后一组下限 8.40、上限 8.599 和组中的代表值 8.40 输入完毕并添加到 Old→New 框中为止。

④ 单击 Continue 按钮，返回 Recode into Different Variables 对话框，单击 OK 按钮，返回数据编辑窗，转换值在组下限中生成。

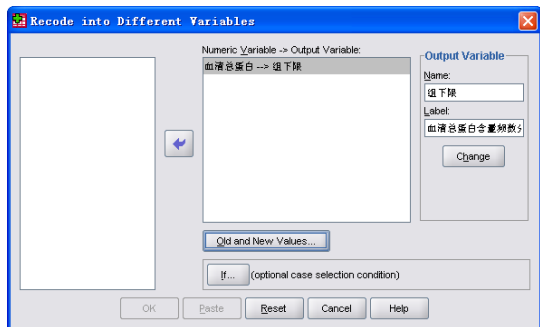


图 2-55 Recode into Different Variables 对话框

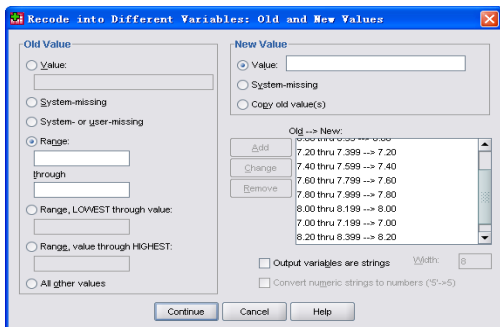


图 2-56 Old and New Values 对话框

5. 制作频数分布表

按 Analyze→Descriptive Statistics→Frequencies 顺序打开 Frequencies 对话框，见图 2-50。选中血清总蛋白含量频数分布表，并将其移入 Variable(s) 的下框中。单击 OK 按钮运行，在输出窗口中得到频数分布表，见表 2-16。

频数分布表从左向右依次列出了组限、组内频数、组内频数百分比、有效值百分比和各组累计百分比。

表 2-16 Frequencies 的运算结果

血清总蛋白含量频数分布表

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 6.4	3	3.0	3.0	3.0
6.6	5	5.0	5.0	8.0
6.8	8	8.0	8.0	16.0
7	13	13.0	13.0	29.0
7.2	25	25.0	25.0	54.0
7.4	23	23.0	23.0	77.0
7.6	9	9.0	9.0	86.0
7.8	7	7.0	7.0	93.0
8	6	6.0	6.0	99.0
8.4	1	1.0	1.0	100.0
Total	100	100.0	100.0	

2.3 数据资料的描述

2.3.1 统计图

统计图是用点、线、面、体来形象地表达数量资料的一种方式，常用的统计图有条图、线图、面积图、圆图、高低图、直方图、序列图、散布图等。

统计图的种类很多，应根据统计资料的类型和目的选用合适的统计图。一般地，定性资料可选用的统计图有条形图、圆图、面积图等；定量资料可选用的统计图有直方图（或多边图）、线图、散点图等。

不同的统计图以不同的方式或姿态来形象化地表达资料。在具体选图时要根据各种统计图的特征来正确选用统计图。

2.3.1.1 条形图

条形图用来表示各相互独立的统计指标的数量大小。通常用纵轴表达数量，用横轴表达分组标志。数量可用绝对数或相对数表达，其数量大小用图中各长条的高度来反映。

1. 分类计数资料条形图

例 2.15 在对北京市 5 所体校运动员的问卷调查研究的数据文件 data02-01.sav 中，用条形图表示 5 所学校参与调查的男、女生人数。

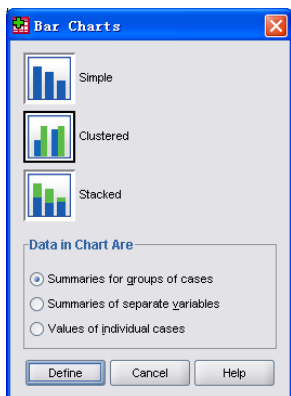


图 2-57 图形菜单

(1) 在 SPSS 数据编辑窗口中，打开 data02-01.sav。

(2) 按 Graphs→Legacy Dialogs→Bar 顺序，打开 Bar 对话框，见图 2-57。

(3) 单击 Clustered，选择有两条或两条以上条组成一组的条形图。单击 Define 按钮（或双击 Clustered），打开 Define Clustered Bars 对话框，见图 2-58。

在左侧的变量名源框中，选择所在学校，单击中间第二个右移按钮，将所在学校送入 Category Axis 框，即定义所在学校为横轴分类轴变量。选择性别变量，用同样的做法，将其送入 Define Clustered By 框中。

(4) 单击 Titles 按钮，打开 Titles 对话框。见图 2-59。在 Line 1 框中，输入“男、女生被调查人数条形图”。单击 Continue 按钮，返回 Define Clustered Bars 对话框。

(5) 其他保持系统默认值，单击 OK 按钮运行，在输出窗口中，得到按要求所做的条形图，见图 2-60。

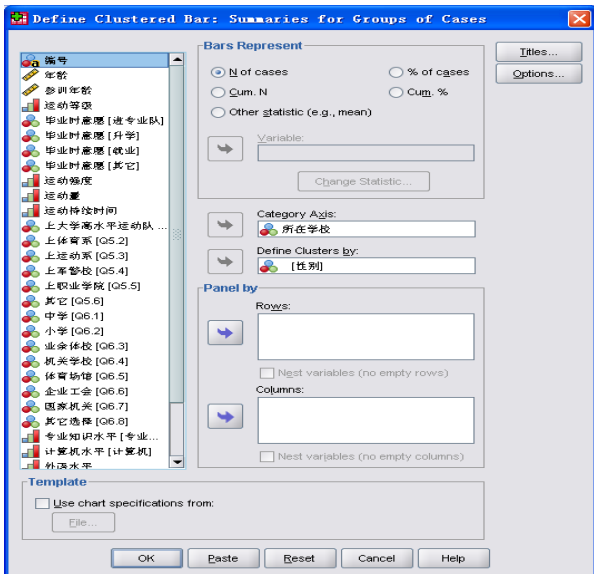


图 2-58 Define Clustered Bars 对话框

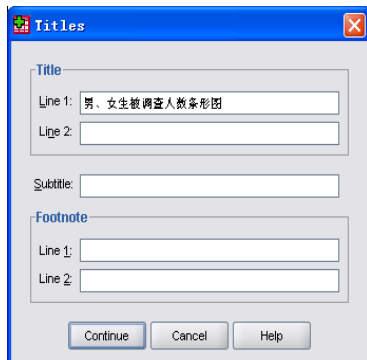


图 2-59 Titles 对话框

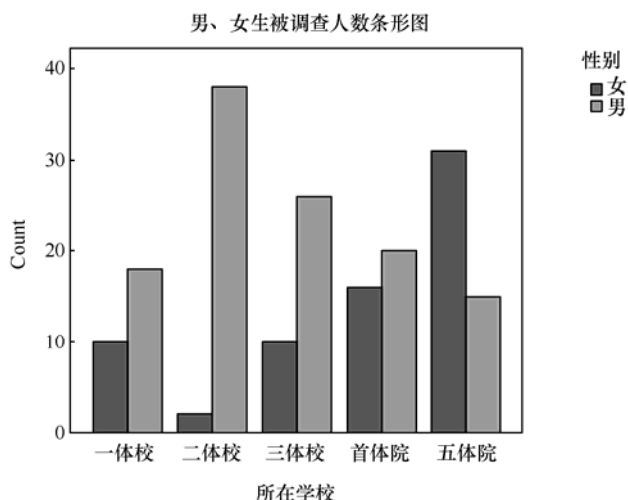


图 2-60 条形图

2. 计量资料条形图

例 2.16 1992 年美国弗吉尼亚《每日评论》报道了过去 50 年海湾海产品收成方面的数据，其中蓝鱼和蓝蟹的观测数据见表 2-17。以此制作条形图。

表 2-17 1940-1990 年海湾的水产收成情况

年代	蓝鱼	蓝蟹	年代	蓝鱼	蓝蟹
1940	15000.0	100000.0	1970	290000.0	4400000.0
1945	150000.0	850000.0	1975	650000.0	4660000.0
1950	250000.0	1330000.0	1980	1200000.0	4800000.0
1955	275000.0	2500000.0	1985	1500000.0	4420000.0
1960	270000.0	3000000.0	1990	2750000.0	5000000.0
1965	280000.0	3700000.0			

制作本例条形图的操作步骤如下：

(1) 在 SPSS 数据编辑窗口中，建立三个变量年代、蓝鱼、蓝蟹，都为数值型尺度变量，用来储存表 2-17 中的各列数据，数据文件见 data02-19.sav。

(2) 按 Graphs→Legacy Dialogs→Bar 顺序，打开 Bar 对话框，见图 2-57。

因为本例中有两个变量要在图形中显示，故在 Data in Chart Are 中选择 Summaries of Separate Variables，对应每个变量生成一个图形，即一个条代表一个变量。

(3) 双击 Simple 样图，打开 Define Simple Bars 对话框，见图 2-61。

在左侧变量名源框中，选定年代，并将其用最下面的右移箭头按钮移入 Column 框中，

选定蓝鱼和蓝蟹，用同样的方法，单击最上面的右移箭头将其移入 **Bars Represent** 框中。在系统默认情况下，自动计算计量资料的平均值，如果要改变这个统计量，可选中需要修改的变量（或组），单击 **Change Statistics** 按钮，即出现 **Statistics** 对话框，见图 2-62。在本例中，由于各对应年代只有一个数据，所以用平均值和用数据值总和一样都等于原始观测值（当然用最大值、最小值也一样）。其他统计量在本例中不合适。故保持系统默认选项，单击 **Continue** 按钮返回 **Define Simple Bars** 对话框。

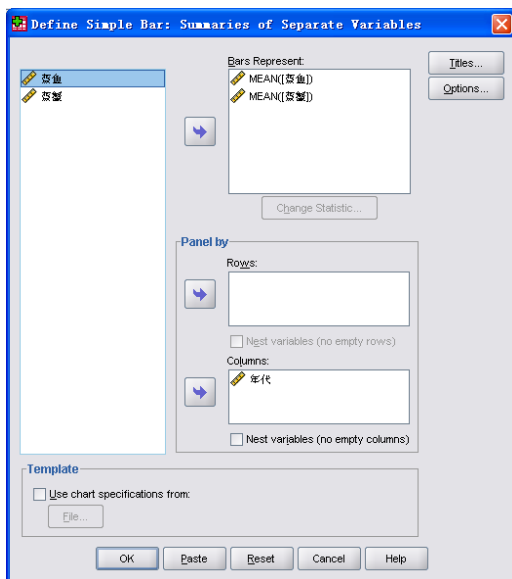


图 2-61 Define Simple Bars 对话框

(4) 单击 **OK** 按钮运行，在输出窗口中得到计量资料的条形图，见图 2-63。

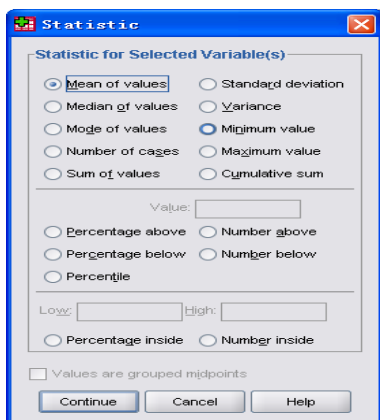


图 2-62 Statistics 对话框

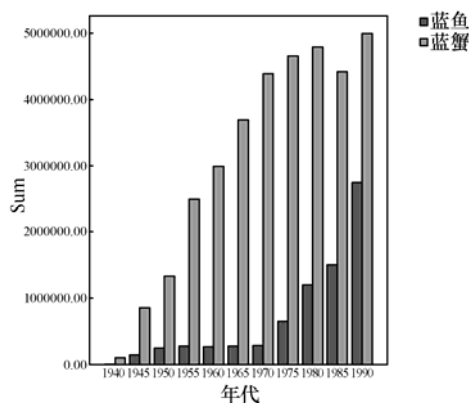


图 2-63 蓝鱼和蓝蟹的条形图

3. 分类计量资料条形图

例 2.17 测得 12 名正常人和 15 名病毒性肝炎患者血清转铁蛋白含量如下：

正常人： 265.4 271.5 284.6 291.3 254.8 275.9 281.7 268.6 264.1 273.2 270.8 260.5

病毒性肝炎患者： 235.9 215.4 251.8 224.7 228.3 231.1 253.0 218.8 233.8 230.9 240.7 221.7 256.9 260.7 224.4

用上述资料制作统计量为平均值的条形图。

(1) 在 SPSS 数据编辑窗口中，设置两个变量，*组别*（数值型名义变量，其值 1=正常人，2=病毒性肝炎患者）和*血清转铁蛋白含量*（数值型尺度变量），录入题中原始数据，并保存在 data02-19-1.sav 数据文件中。

(2) 按 Graphs→Legacy Dialogs→Bar 顺序，打开 Bar 对话框，见图 2-57。

在 Bars Represent 选项中，选择 Other Statistic(e.g.,mean)。由于本例统计量要用平均数，正好是系统默认值，故不必再单击 Change Statistics 按钮去做统计量的选择。选定*组别*，单击第二个右移箭头按钮将其移入 Category Axis 框。

单击 OK 按钮运行，在输出窗口中得到分类计量资料的条形图，见图 2-64。

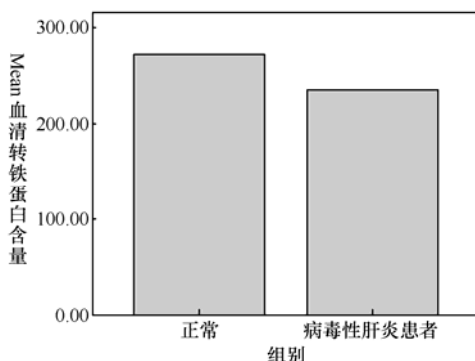


图 2-64 分类计量资料条形图

2.3.1.2 圆图

圆图又称饼图，用来表示事物内部的构成情况。必须使用相对数，且各项之和为 100%。图中各扇形面积表示数量的大小，将 360 度圆心角看成 100%，把每一部分所占的百分数折算成圆心角的度数，根据圆心角的度数就可画出代表各部分数量大小的扇形来。

例 2.18 在对北京市 5 所体校运动员的问卷调查研究的数据文件 data02-01.sav 中，用圆图表示参与调查的男、女生中不同等级运动员对毕业后选择出路方面的分布情况。

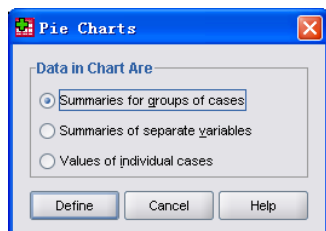


图 2-65 Pie 对话框

(1) 在 SPSS 数据编辑窗口中，打开 data02-01.sav。

(2) 按 Graphs→Legacy Dialogs→Pie 顺序，打开 Pie 对话框，见图 2-65。在 Data in Chart Are 中，选择 Summaries for groups of cases。单击 Define 按钮，打开 Define Pie 对话框，见图 2-66。

在左侧变量名源框中，选择*所在学校*变量，并按第二个右移箭头按钮，将其移入 Define Slices by 框中，用同样的做法，选择*性别*，将其移入 Panel by 下的 Rows

框中,选择运动等级,将其移入 Panel by 下的 Columns 框中。其他保持系统默认选项。

(3) 单击 OK 按钮运行,在输出窗口中得到分类计量资料的饼图,见图 2-67。

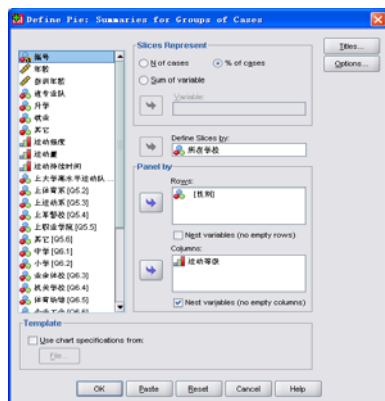


图 2-66 Define Pie 对话框

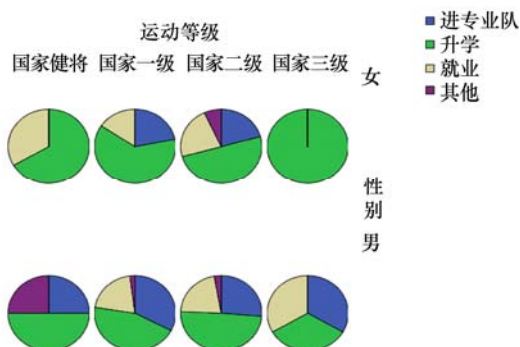


图 2-67 分类计量资料的饼图

2.3.1.3 面积图

面积图用来表示事物的数量在地域上的情况。如反映疾病的地区分布情况、各地的费用开支、劳动成本、高考录取分数线等。

例 2.19 利用表 2-18 中的数据,制作各区劳动费用的面积图。

表 2-18 俄罗斯联邦各区 1958 年谷物生产的劳动费用

各区名称	国营农场	集体农庄
北方区	1.8	5.1
西北区	1.2	3.9
中央区	0.8	2
维亚特区	0.5	2.1
中央黑土	0.4	0.9
伏尔加区	0.2	0.6
北高加索区	0.2	0.4
乌拉尔区	0.3	1.1
西西伯利亚区	0.3	0.4
东西伯利亚区	0.3	0.6
远东区	0.2	0.7

(1) 在 SPSS 数据编辑窗口中, 建立三个变量: 地区 (字符型名义变量)、国营农场费用和集体农庄费用 (数值型尺度变量)。全部数据被录入 data02-20.sav 的数据文件中。

(2) 按 Graphs→Legacy Dialogs→Area 顺序, 打开 Area Charts 对话框, 见图 2-68。在 Data in Chart Are 中, 选择 Summaries of Separate Variables。单击 Stacked 图样, 单击 Define 按钮, 打开 Define Stacked Areas 对话框, 见图 2-69。

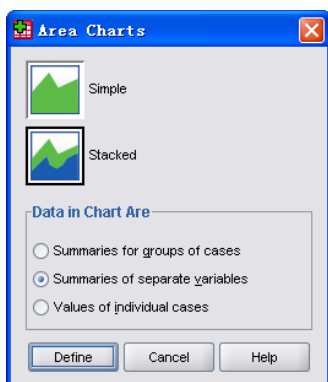


图 2-68 Area Charts 对话框

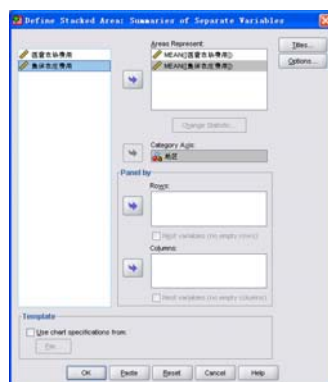


图 2-69 Define Stacked Areas 对话框

在左侧变量名源框中, 选定地区, 并将其用中间第二个右移箭头按钮移入 Category Axis 框中, 选定国营农场费用和集体农庄费用, 用同样的方法, 单击最上面的右移箭头将其移入 Areas Represent 框中。在系统默认情况下, 自动计算计量资料的平均值, 由于每个地区在一个指标上只有一个数据, 所以不用改变这个统计量。其他可保持系统默认值。

(3) 单击 OK 按钮运行, 在输出窗口中得到各地区的计量资料的面积图, 见图 2-70。

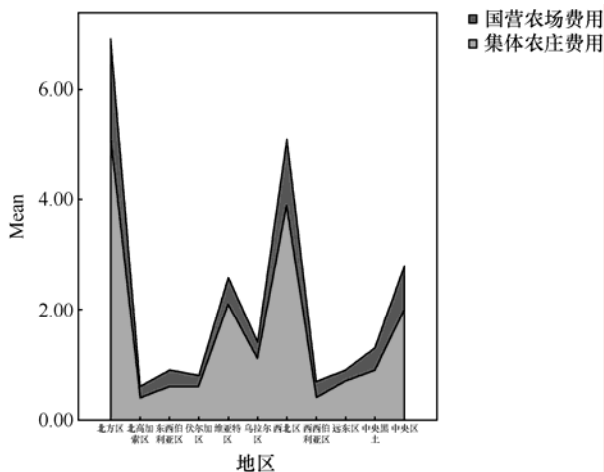


图 2-70 各地区的计量资料的面积图

2.3.1.4 直方图

直方图被用来表示计量资料各组段上频数(或频率)的情况。一般以组限作为横坐标,次数为纵坐标。图中各长条的面积表示各组数量的大小。

1. 用原始计量资料制作直方图

例 2.20 实测 120 名健康成年女子血清蛋白含量(单位: g%) 资料, 见表 2-19。用它来制作血清蛋白含量的直方图。

表 2-19 120 名健康成年女子血清蛋白含量

7.43	7.88	6.88	7.80	7.04	8.05	8.05	6.97	7.12	7.35
7.95	7.56	7.50	7.88	7.20	7.20	7.20	7.43	7.12	7.20
7.50	7.35	7.88	7.43	7.58	6.50	7.43	7.12	6.97	6.80
7.35	7.50	7.20	6.43	7.58	8.03	6.97	7.43	7.35	7.35
7.58	7.58	6.88	7.65	7.04	7.12	8.12	7.50	7.04	6.80
7.04	7.20	7.65	7.43	7.65	7.76	6.73	7.20	7.50	7.43
7.35	7.95	7.35	7.47	6.50	7.65	8.16	7.54	7.27	7.27
6.72	7.65	7.27	7.04	7.72	6.88	6.63	6.73	6.73	7.27
7.58	7.35	7.50	7.27	7.35	7.35	7.27	8.16	7.03	7.43
7.35	7.50	7.20	6.43	7.58	8.03	6.97	7.43	7.35	7.35
7.58	7.58	6.88	7.65	7.04	7.12	8.12	7.50	7.04	6.80
7.35	7.95	7.04	7.15	7.27	7.72	8.43	7.50	7.65	7.04

(1) 在 SPSS 数据编辑窗口中, 建立 *血清蛋白含量* (数值型尺度变量)。全部数据被录入 data02-21.sav 的数据文件中。

(2) 按 Graphs→Legacy Dialogs→Histogram 顺序, 打开 Histogram 对话框, 见图 2-71。选择左侧变量名源框中 *血清蛋白含量*, 单击最上面的右移箭头按钮, 将其移入 Variable 框中, 选择 Display normal curve 选项, 要求在输出直方图时同时画出正态曲线。由于标题一般在后期加工时才加入, 所以, 将对话框中的其他选项保持系统默认状态, 不做修改。

(3) 单击 OK 按钮运行, 在输出窗口中得到血清蛋白含量的直方图和正态曲线图, 见图 2-73。

2. 用频数分布资料制作直方图

例 2.21 从某厂生产的铆钉中随机抽取 200 个测试其直径(单位: mm), 并将其制作成表 2-20 的频数分布表。试据此制作直方图。

(1) 在 SPSS 数据编辑窗口中, 建立两个变量 *组限*、*频数* (数值型尺度变量)。全部数据被录入 data02-22.sav 的数据文件中。

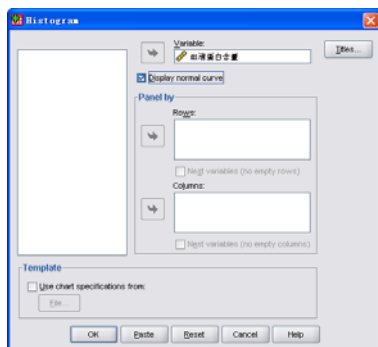


图 2-71 Histogram 对话框

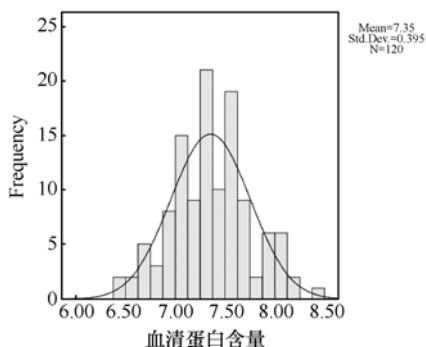


图 2-72 血清蛋白含量的直方图和正态曲线图

表 2-20 频数分布表

组限	13.12	13.16	13.2	13.24	13.28	13.32	13.36	13.4	13.44	13.48	13.52	13.56	13.6	13.64
频数	2	5	10	14	23	22	33	30	19	19	11	7	4	1

(2) 按 Data→Weight Cases 顺序, 打开 Weight Cases 对话框, 见图 2-73。选择 Weight Cases by 选项, 在左侧变量名源框中, 选中 频数变量, 并单击右移箭头将 频数移入 Frequency Variable 框中, 作为加权变量。单击 OK 按钮, 完成加权设置。

(3) 按 Graphs→Legacy Dialogs→Histogram 顺序, 打开 Histogram 对话框, 见图 2-71。选择左侧变量名源框中“组限”, 单击最上面的右移箭头按钮, 将其移入 Variable 框中, 选择 Display normal curve 选项, 要求在输出直方图时同时画出正态曲线。对话框中的其他选项保持系统默认状态。

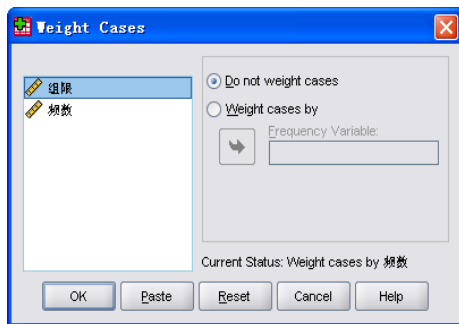


图 2-73 Weight Cases 对话框

(4) 单击 OK 按钮运行, 在输出窗口中得到 200 个铆钉直径数据的直方图和正态曲线图, 见图 2-74。

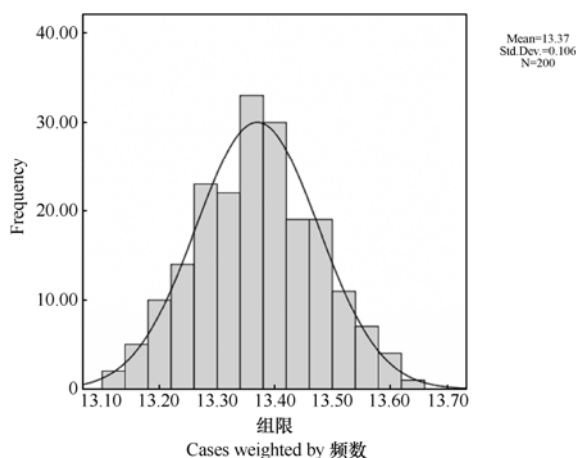


图 2-74 200 个铆钉直径数据的直方图和正态曲线图

2.3.1.5 线图

所谓线图是指在直角坐标系中，用线段的升降表示事物或现象随时间或其他指标发展变化的统计图。线图一般适用于包含着 2 个计量指标的资料。在横轴上的计量指标通常是时间，在纵轴上的计量指标通常是某种率。纵、横轴上的尺度一律用算术尺度。它适合于表达 1 个或多个事物或现象随着时间的推移，数量的增减幅度。

例 2.22 以例 2.16 中蓝鱼、蓝蟹数据为例，制作线图。

(1) 在 SPSS 数据编辑窗口中，打开 data02-19.sav。

(2) 按 Graphs→Legacy Dialogs→Line 顺序，打开 Line Charts 对话框，见图 2-75。单击 Multiple 样图，选多重线图。单击 Define 按钮，打开 Define Multiple Line 对话框，见图 2-76。

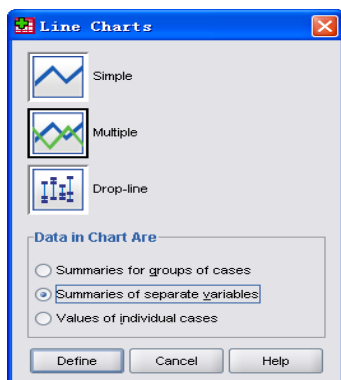


图 2-75 Line Charts 对话框

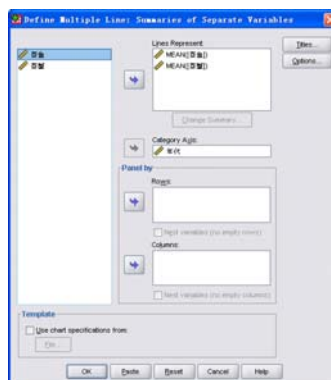


图 2-76 Define Multiple Line 对话框

选择年代,单击中间第二个右移箭头按钮,将其移入 Category Axis 框中,设定年代作为横轴变量,选择蓝鱼、蓝蟹,单击中间第一个右移箭头按钮,将它们移入 Lines Represent 框中。对话框中的其他选项保持系统默认状态。

(3) 单击 OK 按钮运行,在输出窗口中得到蓝鱼、蓝蟹 50 年间随年代变化的线图,见图 2-77。

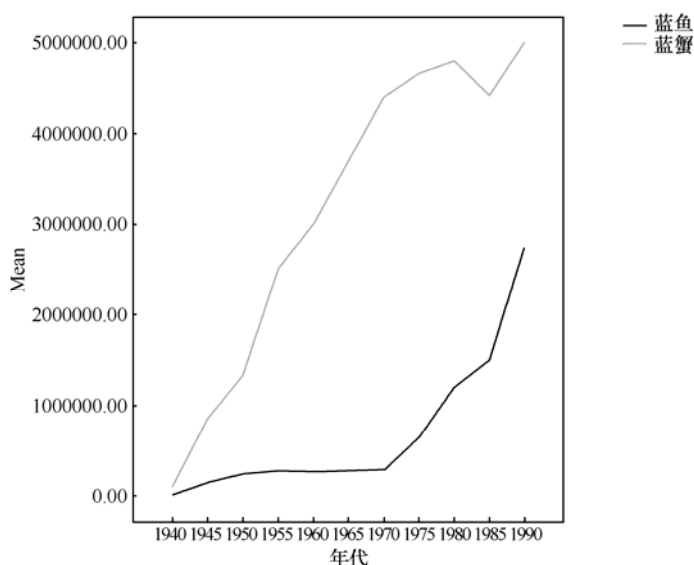


图 2-77 蓝鱼、蓝蟹随年代变化的线图

2.3.1.6 散点图

散点图一般适用于含有两个或两个以上计量指标的数据资料。如果两变量之间有自变量与因变量之分时,通常把自变量放在横轴上,把因变量放在纵轴上。将成对的数据(X , Y)在直角坐标系中用圆点表示出来就称为散点图或散布图。它可以形象地反映出在专业上有一定联系的两个连续变量之间的变化趋势,因此,可借助它来帮助判断是否值得进行直线相关和回归分析或拟合何种类型的曲线方程。

例 2.23 测得某大学二年级 12 名女生的体重与肺活量数据,见表 2-21。试制作散点图。

表 2-21 12 名女生的体重与肺活量数据

编号	1	2	3	4	5	6	7	8	9	10	11	12
体重 (kg)	42	42	46	46	46	50	50	50	52	52	58	58
肺活量 (L)	2.6	2.2	2.8	2.4	2.8	3.4	3.1	3.5	2.9	3.5	3.0	3.0

(1) 在 SPSS 数据编辑窗口中, 建立 *体重*、*肺活量* (数值型尺度变量)。全部数据被录入到 data02-23.sav 的数据文件中。

(2) 单击 **Graphs**→**Legacy Dialogs**→**Scatter /Dot** 顺序, 打开 **Scatter /Dot** 对话框, 见图 2-78。因为本例中讨论的变量只有最基本的两个, 所以, 单击 **Simple Scatter** 样图, 选择简单散点图。

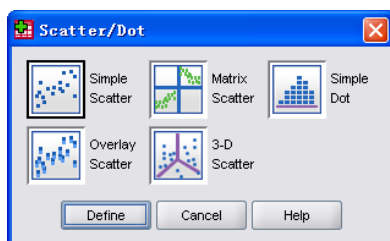


图 2-78 Scatter /Dot 对话框

(3) 单击 **Define** 按钮, 打开 **Simple Scatterplot** 对话框, 见图 2-79。选择左侧变量名源框中 *肺活量*, 单击最上面的右移箭头按钮, 将其移入 **Y Axis:** 框中, 即确定肺活量为纵轴变量。选择左侧变量名源框中 *体重*, 单击最上面第二个右移箭头按钮, 将其移入 **X Axis:** 框中, 确定体重为横轴变量。对话框中的其他选项保持系统默认状态。

(4) 单击 **OK** 按钮运行, 在输出窗口中得到体重与肺活量的散点图, 见图 2-80。

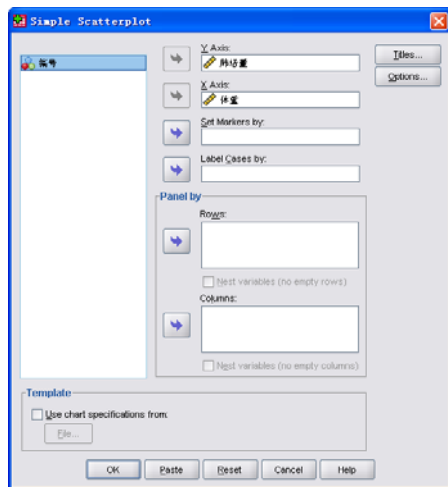


图 2-79 Simple Scatterplot 对话框

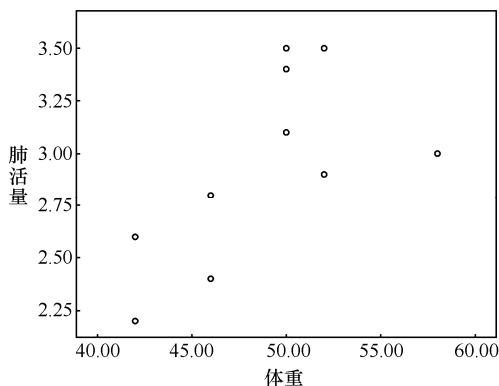


图 2-80 体重与肺活量的散点图

例 2.24 某销售化妆品的公司, 统计了一个月内在 15 个城市的销售情况, 统计指标包括销售量 (箱)、使用该化妆品的人数和他们的人均收入, 统计结果见表 2-22。

表 2-22 15 个城市的销售数据资料

地区	销售量	人数	人均收入
1	162	274000	2450
2	120	180000	3250
3	223	375000	3802
4	131	205000	2838
5	67	86000	2347
6	169	265000	3782
7	81	98000	3008
8	192	330000	2450
9	116	195000	2137
10	55	53000	2560
11	252	430000	4020
12	232	372000	4427
13	144	236000	2660
14	103	157000	2088
15	212	370000	2605

(1) 在 SPSS 数据编辑窗口中, 建立 4 个变量, *地区* (数值型名义变量), *销售量*、*人数*、*人均收入* (数值型尺度变量)。全部数据被录入 data02-24.sav 的数据文件中。

(2) 单击 Graphs→Legacy Dialogs→Scatter /Dot 顺序, 打开 Scatter /Dot 对话框, 见图 2-78。因为本例中讨论的变量有 3 个, 所以, 单击 Matrix Scatter 样图, 打开 Scatterplot Matrix 对话框, 选择简矩阵散点图, 见图 2-81。

在左侧变量名源框中, 选中 *销售量*、*人数*、*人均收入*, 单击中间最上面的右箭头按钮, 将它们移入 Matrix Variables 框中, 其他保持系统默认值。

(3) 单击 OK 按钮运行, 在输出窗口中得到 *销售量*、*人数*、*人均收入* 的散点矩阵图, 见图 2-82。

2.3.1.7 3-D 图

3-D 图 (又称立体图) 是在三位空间中描述三个变量之间关系的图形。*X* 和 *Z* 轴可以都是分类变量, 也可以是单独变量和单个样品, 不能同时都是单独变量或单个样品。

例 2.25 为研究年龄、呼吸情况、吸烟与肺癌之间的关系, 研究人员共调查了年龄小于等于 40 岁的男性 1483 人, 年龄大于 40 岁的男人 1178 人, 得到如下的基本情况资料, 见表 2-23。试用 3-D 图描述这些基本情况。

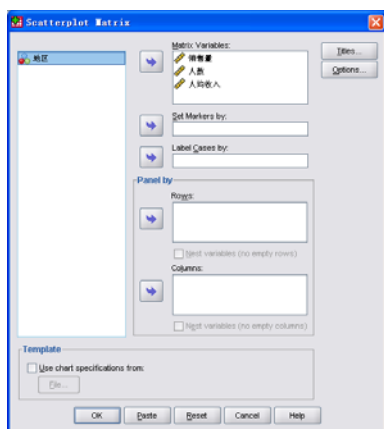


图 2-81 Scatterplot Matrix 对话框

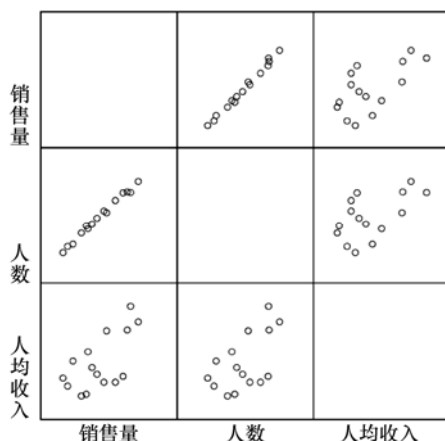


图 2-82 销售量、人数、人均收入的散点矩阵图

表 2-23 被调查者的基本情况构成

		不吸烟	吸烟
年龄≤40	呼吸正常	567	874
	呼吸不正常	14	28
年龄>40	呼吸正常	328	780
	呼吸不正常	2	68

(1) 在 SPSS 数据编辑窗口中, 建立 4 个变量, 分别是年龄、呼吸、吸烟情况 (数值型名义二分变量) 和观察值 (数值型尺度变量), 并录入表 2-23 中数据, 存储在数据文件 data02-25.sav 中。

(2) 单击 Data→Weight Cases 顺序, 打开 Weight Cases 对话框, 见图 2-73。选择 Weight Cases by 选项, 在左侧变量名源框中, 选中观察值变量, 并单击右移箭头将频数移入 Frequency Variable 框中, 作为加权变量。单击 OK 按钮, 完成加权设置。

(3) 单击 Graphs→Legacy Dialogs→3-D Bars 顺序, 打开 3D Bar Charts 对话框, 见图 2-83。在 X-axis represents 选项中选择 Groups of cases, 选用样品分类模式。在 Z-axis represents 选项中也选择 Groups of cases, 表明本例中的 X、Z 轴上都放置分类变量。

(4) 单击 Define 按钮, 打开 Define 3-D Bar 对话框, 见图 2-84。

在左侧变量名源框中, 选择年龄, 单击中间第二个右移箭头按钮, 将其移入 X Category Axis 框中, 选择呼吸, 单击中间第三个右移箭头按钮, 将其移入 Z Category Axis 框中, 选择吸烟情况, 单击右侧第一个右移箭头按钮, 将其移入 Stack/Cluster by (up to 2 Variables) 下的 Stack 框中。要求对吸烟和不吸烟作堆处理。

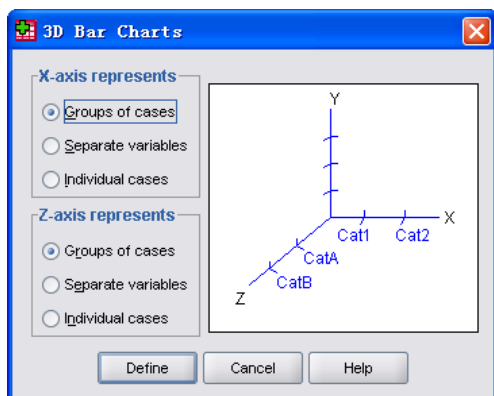


图 2-83 3D Bar Charts 对话框

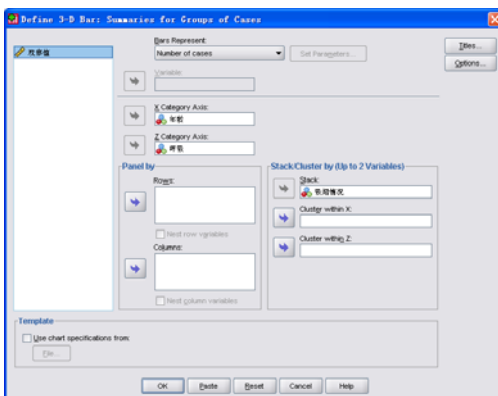


图 2-84 Define 3-D Bar 对话框

(5) 单击 OK 按钮，在输出窗口中得到 3-D 图的输出结果，见图 2-85。

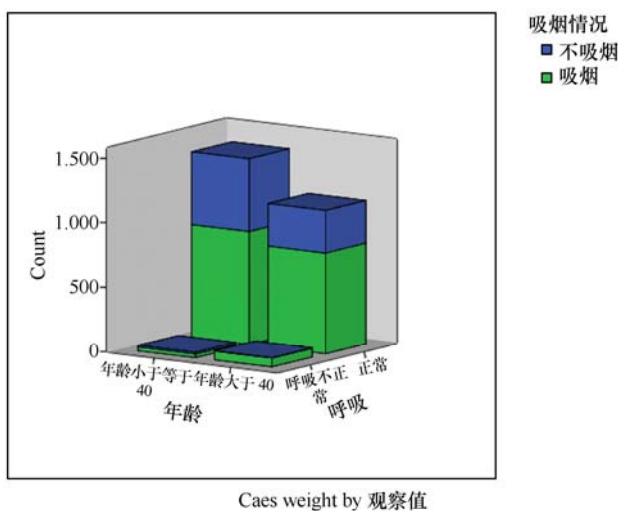


图 2-85 被调查人员的基本情况

2.3.1.8 高低图

高低图是一种标志某些现象在单位时间内变化情况的统计图。既可描述某些现象在短期内的波动，也可说明它们长期变化的趋势。

例 2.26 在 data02-26.sav 中，记录了岷江水电在 2008 年 11 月 5 日至 11 月 19 日之间的开盘价、最高价、最低价和收盘价的交易情况，试作高低图。

(1) 在 SPSS 数据编辑窗口中，打开数据文件 data02-26.sav。

(2) 按 Graphs→Legacy Dialogs→High-Low 顺序, 打开 High-Low Charts 对话框, 见图 2-86。由于我们关心的是股票的最高价、最低价和收盘价, 因此, 单击 Simple High-Low-Close 样图, 选择简单的高-低-收盘图。在 Data in Chart Are 选项中, 选择 Values of individual cases 选项。

(3) 单击 Define 按钮, 打开 Define Simple High-Low-Close 对话框, 见图 2-87。

在左侧变量名源框中, 选择 岷江水电最高股价, 单击中间第一个右移箭头按钮, 将其移入 High 下框中; 选择 岷江水电最低股价, 按中间第二个右移箭头按钮, 将其移入 Low 下框中; 选择 岷江水电收盘价, 按中间第三个右移箭头按钮, 将其移入 Close 下框中。在 Category Labels 选项中, 选择 Variable, 在左侧变量名源框中, 选择 时间, 按 Variable 选项下的右移箭头按钮, 将 时间 移入右侧的框中。

(4) 单击 OK 按钮, 在输出窗口中得到 Simple High-Low-Close 图的输出结果, 见图 2-88。

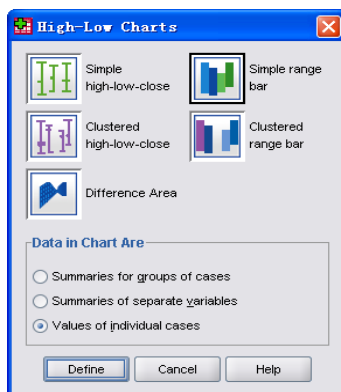


图 2-86 High-Low Charts 对话框



图 2-87 Define High-Low Charts 对话框

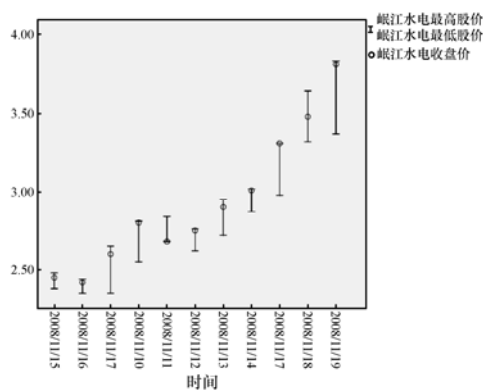


图 2-88 Simple High-Low-Close 图

2.3.2 统计表

表格是用来简明地表达事物间数量关系的一种方式。统计表格主要由表号、标题、线条、分组标志、观测指标、文字和数字组成。

制表的原则是要根据具体内容合理安排好分组标志和观测指标。制表中要力求简单明了, 中心突出。

统计表格一般分简单表和复杂表两种。只含 1 个分组标志的表称为简单表，含 2 个及 2 个以上分组标志的表称为复合表。

2.3.2.1 制作列联表

当分组标志为定性资料时，观测指标一般为频数和百分比(率)；如果表中的原始数据只有 2 个分组标志，且每个分组标志只有 2 个水平时，称为四格表或 2×2 表，如果 2 个分组标志分别有 R 个和 C 个水平时，称为 $R \times C$ 表。

例 2.27 用使君子治疗 184 例蛔虫、蛲虫混合感染患者的排虫情况，排出蛔虫 56 例、蛲虫 36 例，既排出蛲虫又排出蛔虫的 16 例，都没有排出的 76 例。请根据这些数据资料制作列联表。

在 SPSS 中操作步骤如下：

(1) 在数据编辑窗口中进行变量设计

在数据文件中，建立三个变量，即蛔虫变量与蛲虫变量：数值型、名义变量，值标签为：1=排出、2=未排出；频数变量：数值型、尺度变量。

(2) 录入原始数据

在数据编辑窗口中建立如图 2-89 所示的数据文件（见 data02-27.sav）。

(3) 按 Data→Weight Cases 顺序，打开 Weight Cases 对话框，见图 2-73。选择 Weight Cases by 选项，在左侧变量名源框中，选中 频数变量，并单击右移箭头将 频数移入 Frequency Variable 框中，作为加权变量。单击 OK 按钮，完成加权设置。

(4) 按 Analyze→Descriptive Statistics→Crosstabs，展开 Crosstabs 对话框，见图 2-90。在左侧变量名源框中，选择 蛔虫变量，单击最上面的右移箭头将其移入 Row(s)框中，选择 蛲虫变量，单击中间的右移箭头将其移入 Column(s)框中。

	蛔虫	蛲虫	频数
1	1.00	1.00	16.00
2	1.00	2.00	56.00
3	2.00	1.00	36.00
4	2.00	2.00	76.00

图 2-89 数据文件格式

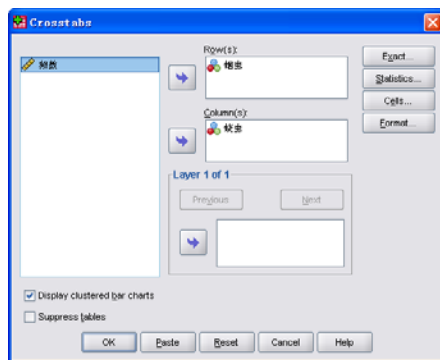


图 2-90 Crosstabs 对话框

(5) 单击 Cells 按钮，打开 Cells 对话框，见图 2-91。在 Counts 选项中选择 Observed 选项，要求输出观察数。在 Percentages 选项中选择 Row、Column、Total，要求输出行、

列和总百分比。其他保持系统默认值。单击 Continue 按钮返回 Crosstabs 对话框。

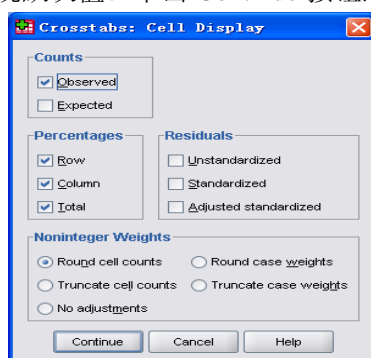


图 2-91 Cell 对话框

(6) 单击 OK 按钮运行，在输出窗口中得到输出表格，见表 2-24。

表 2-24 蛔虫与蛲虫列联表资料

蛔虫 • 蛲虫 Crosstabulation					
			蛲虫		Total
			排出	未排出	
蛔虫	排出	Count	16	56	72
		% within 蛔虫	22.2%	77.8%	100.0%
		% within 蛲虫	30.8%	42.4%	39.1%
		% of Total	8.7%	30.4%	39.1%
	未排出	Count	36	76	112
		% within 蛔虫	32.1%	67.9%	100.0%
		% within 蛲虫	69.2%	57.6%	60.9%
		% of Total	19.6%	41.3%	60.9%
Total		Count	52	132	184
		% within 蛔虫	28.3%	71.7%	100.0%
		% within 蛲虫	100.0%	100.0%	100.0%
		% of Total	28.3%	71.7%	100.0%

2.3.2.2 制作多重应答表

在调查问卷设计中，有时研究者会将性质相近的素质、能力或其他方面的因素等集中在一起，组成一个综合问题，让被调查者在一个给定范围的相同选项上给出评判或做出选择，这相当于将多个单选题集中在一起，组合成一个多重单选题。显然能否组成一个多重单选题的关键在于是否能组成相同的被选项。

现对此类问题的被调查者的选择结果制作多重应答表。

例 2.28 在对北京市 5 所体校运动员的问卷调查研究中, 问题 7 “请在下述各项上, 给自己的水平和能力作如实的评价 (在选中项上打√)”:

	很好	较好	一般	不好	很不好
专业知识水平					
计算机水平					
英语水平					
语言表达能力					
示范能力					
教学训练能力					

在 SPSS 中, 已将上述 6 个知识与能力方面的问题设置成 6 个变量, 参见第 2 章 2.1.2.4 中的变量定义, 具体调查数据见 data02-01.sav 中相应的变量中所列。

在 SPSS 中的操作步骤如下:

(1) 在 SPSS 数据编辑窗口中, 打开数据文件 data02-01.sav。

(2) 按 Analyze→Tables→Custom Tables 顺序, 打开 Custom Tables 对话框, 见图 2-92。

(3) 在变量名源框中, 选定专业知识水平、计算机水平、英语水平、语言表达能力、示范讲解、教学训练能力, 并将其拖曳到探究窗口的 Rows 上, 在 Category Position 选项中, 选择 Row Labels in Columns, 按列列出行标签。其他保持系统默认选择。

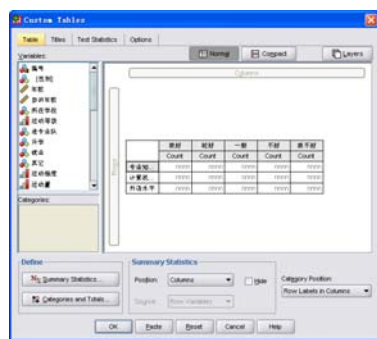


图 2-92 Custom Tables 对话框

(4) 单击 OK 按钮运行, 在输出窗口中出现运行结果, 见表 2-25。

表 2-25 多重应答调查结果表

	很好	较好	一般	不好	很不好
	Count	Count	Count	Count	Count
专业知识水平	40	86	56	3	1
计算机水平	11	35	111	23	6
外语水平	3	10	68	84	21
语言表达能力	34	58	80	13	1
示范讲解	23	52	88	18	5
教学训练能力	35	81	60	7	3

2.3.2.3 制作嵌套表

嵌套表是由二张或多张列联表合并而成的一张复合表。

例 2.29 在对北京市 5 所体校运动员的问卷调查研究的数据文件 data02-01.sav 中，列出 5 所学校参与调查的男、女生的运动等级方面的基本情况。

(1) 在 SPSS 数据编辑窗口中，打开 data02-01.sav。

(2) 按 Analyze→Tables→Custom Tables 顺序，打开 Custom Tables 对话框（图 2-92）。

(3) 在变量名源框中，选定所在学校，并将其拖曳到探究窗口的 Rows 上，选定性别，将其拖曳到的所在学校下一层，选定运动等级并将其拖曳到探究窗口的 Columns 上。其他保持系统默认选项。

(4) 单击 OK 按钮运行，在输出窗口中出现运行结果，见表 2-26。

表 2-26 5 所学校参与调查学生的基本情况表

			运动等级				
			国家健将	国家一级	国家二级	国家三级	无级
			Count	Count	Count	Count	Count
所在学校	一体校	女	0	3	7	0	0
		男	1	1	16	0	0
	二体校	女	0	1	1	0	0
		男	0	4	33	1	0
	三体校	女	0	3	7	0	0
		男	0	10	16	0	0
	首体院	女	1	2	11	2	0
		男	0	2	17	1	0
	五体校	女	5	15	11	0	0
		男	2	10	3	0	0

2.3.3 统计量

2.3.3.1 集中性统计量

1. 均值 (Mean)

(1) 意义：表示变量所有取值的集中趋势或平均水平。

(2) 它是用所有观测值的和除以观测次数来计算。计算公式为

$$\text{总体均值: } \mu = \sum_{i=1}^N X_i / N, \quad \text{样本均值: } \bar{x} = \sum_{i=1}^n x_i / n$$

若一组样本观察值 x_1, x_2, \dots, x_n ，分别出现了 f_1, f_2, \dots, f_n 次，则定义 $\bar{x} = \frac{\sum_{i=1}^n x_i f_i}{\sum_{i=1}^n f_i}$ 为样本

加权算术平均数。

(3) 应用时的注意事项：当数据资料服从或近似服从正态分布时，均值是描述集中趋势的理想指标。

(4) 实例

例 2.30 某班 20 名学生在统计学期末考试中的成绩分别为：90, 88, 78, 67, 50, 78, 82, 94, 68, 82, 79, 76, 79, 59, 74, 81, 86, 78, 85, 88 (单位：分)，求该班的平均成绩。数据已存放在 data02-28.sav 数据文件中。

在 SPSS 中的解题步骤：

① 按 Analyze→Descriptive Statistics→Frequencies 顺序，展开 Frequencies 对话框，见图 2-50。在变量名表中，选择所要分析的考试成绩变量，并将其移到 Variables 矩形框中。

② 单击 Statistics 按钮，弹出 Statistics 对话框，见图 2-93。

在 Central Tendency (集中趋势) 下选择 Mean 选项，要求计算样本均值。

单击 Continue 按钮，返回 Frequencies 对话框。

其他保持系统默认选项。

③ 单击 OK 按钮运行，则在输出窗口中得到均值计算表，见表 2-27。

可见，该班学生在统计学期末考试中的平均成绩为 78.1 分。

例 2.31 现有 125 名女生跳远成绩的频数分布表资料如表 2-28。求该批连续型频数分布表数据资料的加权算术平均数。在 SPSS 中，处理本类问题时的具体步骤如下：

表 2-27 均值表

Statistics		
考试成绩		
N	Valid	20
	Missing	0
Mean		78.1000

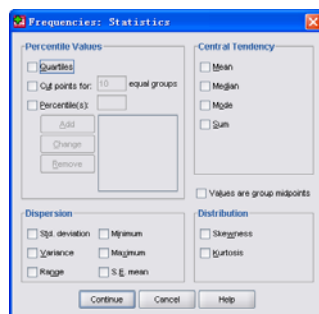


图 2-93 Statistics 对话框

表 2-28 125 名女生跳远成绩的频数分布表 (单位：米)

分组	3.80~	4.00~	4.20~	4.40~	4.60~	4.80~	5.00~	5.20~
频数	9	7	11	28	26	21	16	7

① 在 SPSS 中，建立数据文件，见 data02-29.sav。

② 加权处理

按 Data→Weight Cases 顺序, 打开 Weight Cases 对话框, 见图 2-73。选择 Weight Cases by 选项, 在左侧变量名源框中, 选中 频数变量, 并单击右移箭头将 频数移入 Frequency Variable 框中, 作为加权变量。单击 OK 按钮, 完成加权设置。

③ 组中代表值的确定

在连续型频数分布表资料中, 各组的代表性观察值通常用各组的组中值来替代, 因此, 在 SPSS 中, 可以用本章 2.2 中的方法来计算组中值的计算。方法如下:

按 Transform→Compute Variable 顺序打开 Compute Variable 对话框, 见图 2-53。在 Target Variable 的下框中输入目标变量名 组中值。在 Numeric Expression 的下框中输入: 组下限+0.1 (因为本例为等组距分组, 所以组中值=组下限+组距/2=组下限+0.1)。

单击 OK 按钮执行, 则在工作的数据文件中出现 组中值 一列新变量及其观察值。见图 2-94。

④ 按 Analyze→Descriptive Statistics→Descriptives 顺序, 展开 Descriptives 对话框, 见图 2-95。

	组下限	频数	组中值
1	3.80	9.00	3.90
2	4.00	7.00	4.10
3	4.20	11.00	4.30
4	4.40	28.00	4.50
5	4.60	26.00	4.70
6	4.80	21.00	4.90
7	5.00	11.00	5.10
8	5.20	7.00	5.30

图 2-94 各组的组中值



图 2-95 Descriptives 对话框

在左侧变量名源框中选择 组中值, 将其移入 Variable(s) 下框中。

⑤ 单击 Options 按钮, 弹出 Options 对话框, 见图 2-96。

在 Options 对话框中, 选择 Mean 选项, 单击 Continue 按钮返回 Descriptives 对话框。

⑥ 单击 OK 按钮执行, 在输出窗口中, 出现计算结果, 见表 2-29。

从表 2-29 可见, 样本含量 N 为 125, 有效样本含量也为 125, 说明全部数据都是有效的。125 名女生跳远成绩的频数分布表资料的算术平均数为 4.6472。

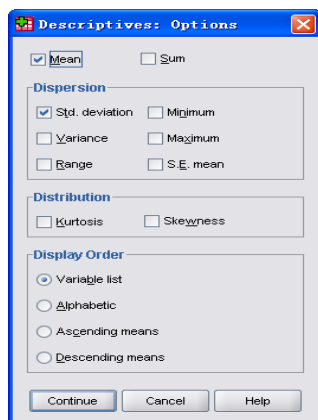


图 2-96 Options 对话框

表 2-29 计算结果

Descriptive Statistics		
	N	Mean
组中值	125	4.6472
Valid N (listwise)	125	

2. 中位数 (Median)

(1) 定义：中位数也是反映集中趋势的指标。它是指处在一组有序数据最中间的数。

(2) 计算：设该组数据共有 N 个，则当该组数据为偶数时，中位数为： $M = x_{N/2}$ ，而当该组数据为奇数时，中位数为： $M = x_{(N+1)/2}$ 。

(3) 应用：当数据资料呈偏态分布时，用中位数做数据资料的集中趋势指标要优于均值。

(4) 实例

例 2.32 仍以例 2.30 为例，数据存放在 data02-28.sav 中，求 20 名学生统计学期末考试成绩的中位数。

① 按 Analyze→Descriptive Statistics→Frequencies 顺序，展开 Frequencies 对话框，见图 2-50。在变量名表中，选择所要分析的考试成绩变量，并将其移到 Variables 矩形框中。

② 单击 Statistics 按钮，弹出 Statistics 对话框，见图 2-93。

在 Central Tendency (集中趋势) 下选择 Median 选项，要求计算样中位数。

单击 Continue 按钮，返回 Frequencies 对话框。

其他保持系统默认选项。

③ 单击 OK 按钮运行，则在输出窗口中得到中位数计算表，见表 2-30。

可见，该班学生在统计学期末考试中的中位数成绩为 79.0000 分。

例 2.33 求例 2.31 中 125 名女生跳远成绩的中位数。

在 SPSS 中，处理本类问题时的具体步骤如下：

① 在 SPSS 中，打开数据文件 data02-29.sav。

② 按例 2.31 中 2. 的做法用频数变量进行加权处理。

③ 按例 2.31 中 3. 的做法确定组中值。

④ 按 Analyze→Descriptive Statistics→Frequencies 顺序，展开 Frequencies 对话框，见图 2-50。

在左侧变量名源框中选择组中值，将其移入 Variable(s) 下框中。

表 2-30 中位数表

Statistics		
考试成绩		
N	Valid	20
	Missing	0
Median		79.0000

⑤ 单击 Statistics 按钮, 弹出 Statistics 对话框, 见图 2-93。

在 Statistics 对话框中, 选择 Median 选项, 单击 Continue 按钮返回 Frequencies 对话框。

⑥ 单击 OK 按钮执行, 在输出窗口中, 出现计算结果, 见表 2-31。

从表 2-31 可见, 样本含量 N 为 125, 缺失值为 0, 说明全部数据都是有效的。125 名女生跳远成绩的频数分布表资料的中位数为 47.0000。

表 2-31 计算结果

Statistics		
组中值		
N	Valid	125
	Missing	0
Median		47.0000

3. 众数 (Mode)

(1) 定义: 众数也是反映集中趋势的指标。它是指样本观察值中出现次数最多的那个观察值。

(2) 应用: 当数据资料呈偏态分布时, 用众数表示不需要做进一步处理的一些间断性变异资料的集中性, 此时用它做数据资料的集中趋势指标要优于均值。

(3) 实例

例 2.34 求例 2.31 中 125 名女生跳远成绩的众数。

在 SPSS 中, 处理本类问题时的具体步骤如下:

① 在 SPSS 中, 打开数据文件 data02-29.sav。

② 按例 2.31 中 2. 的做法用频数变量进行加权处理。

③ 按例 2.31 中 3. 的做法确定组中值。

④ 按 Analyze→Descriptive Statistics→Frequencies 顺序, 展开 Frequencies 对话框, 见图 2-50。

在左侧变量名源框中选择组中值, 将其移入 Variable(s) 下框中。

⑤ 单击 Statistics 按钮, 弹出 Statistics 对话框, 见图 2-93。

在 Statistics 对话框中, 选择 Mode 选项, 单击 Continue 按钮返回 Frequencies 对话框。

⑥ 单击 OK 按钮执行, 在输出窗口中, 出现计算结果, 见表 2-32。

从表 2-32 可见, 样本含量 N 为 125, 缺失值为 0, 说明全部数据都是有效的。125 名女生跳远成绩的频数分布表资料的众数为 4.50。

表 2-32 计算结果

Statistics		
组中值		
N	Valid	125
	Missing	0
Mode		4.50

4. 几何均数 (Geometric mean)

(1) 定义: 几何均数就是随机变量 X 的 n 个观察值乘积的 n 次方根。

(2) 几何均数的计算公式: $G = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$ 。

其中: G —几何均数; x_i —变量第 i 个观察值; n —样本含量

(3) 适用条件

当数据资料成等比数列，服从或近似服从对数正态分布时，可用几何均数表示其集中趋势。如在医学研究中的抗体滴度、某些物质浓度等，其数据特点是变量值间接倍数关系变化；在经济研究中，研究的销售量、成本、价格等的增长率问题都可以用几何均数来描述其集中趋势。

(4) 实例

例 2.35 69 例 RA 患者血清 EBV-VCA-IgG 抗体测定结果见表 2-33。求其几何均数。

表 2-33 69 例 RA 患者血清抗体测定结果

	抗体滴度	滴度倒数	频数
1	1:10	10.00	4.00
2	1:20	20.00	3.00
3	1:40	40.00	10.00
4	1:80	80.00	10.00
5	1:160	160.00	11.00
6	1:320	320.00	15.00
7	1:640	640.00	14.00
8	1:1280	1280.00	2.00

在已知频数分布状态下，用 SPSS 处理类似本例的基本步骤如下：

① 在 SPSS 中，建立数据文件，见 data02-30.sav。

② 对滴度倒数作加权处理。

具体做法同上例，权重变量为频数。

③ 按 Analyze→Reports→Case Summaries 顺序，打开 Case Summaries 对话框，见图 2-51。在左边的变量名源框中，选中滴度倒数，将其移入 Variables 框中，关闭本对话框中的其他选项。

④ 单击 Statistics 按钮，展开 Statistics 对话框，见图 2-52。在 Statistics 下框中选择 Geometric Mean 将其移入 Cell Statistics 框中，要求计算几何均数。单击 Continue 按钮返回 Case Summaries 对话框。

⑤ 单击 OK 按钮执行，在输出窗口中出现计算结果，见表 2-34。

所以，69 例 RA 患者血清 EBV-VCA-IgG 抗体滴度测定结果的几何均数为 1: 150.64。

表 2-34 几何均数

Case Summaries
Geometric Mean
滴度倒数
1.5064E2

5. 调和均数(Harmonic Average)

(1) 定义：调和平均数又称倒数平均数，是变量倒数的算术平均数的倒数。

(2) 计算公式

$$\text{简单调和均数: } H = \frac{1}{(\sum_{i=1}^n \frac{1}{x_i})/n} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

$$\text{加权调和均数: } H = \frac{1}{(\sum_{i=1}^n \frac{1}{x_i} f_i) / \sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n f_i}{\sum_{i=1}^n \frac{1}{x_i} f_i}$$

(3) 适用条件

调和均数 H 适合于表达呈极严重的正偏态资料的平均水平。目测法:多数数据比较接近且数值较小,个别数据特别大。通过取倒数变换后,使个别特别大的数据在均数中起的作用变小,使所求得的均数能更好的反映绝大多数数据的平均水平。

调和平均数易受极端值的影响,且受极小值的影响比受极大值的影响更大;只要有一个变量值为零,就不能计算调和平均数;当频数分布表数列有开口组时,其组中值即使按相邻组距计算了,假定性也很大,这时,调和平均数的代表性就很不可靠;调和平均数应用的范围较小。

(4) 实例

例 2.36 用一定剂量的缓叭妥使 7 只大鼠的睡眠持续时间(分)分别为 25,30,55,50,35,26,>120(最后一只鼠虽等了 2 小时仍不苏醒),求大鼠的平均睡眠时间。

【题析】 由于本例数据明显呈极严重的正偏态,所以,用调和均数来描述其集中趋势。由于最后一只鼠等了 2 小时仍不苏醒,所以其睡眠时间应当作无穷大,在 SPSS 的实际计算中取 120×10^{10} 。

在 SPSS 中,解题步骤如下:

① 在 SPSS 中,建立数据文件,见 data02-31.sav。

② Analyze→Reports→Case Summaries 顺序,打开 Case Summaries 对话框,见图 2-51。在左边的变量名源框中,选中睡眠时间,将其移入 Variables 框中,关闭本对话框中的其他选项。

③ 单击 Statistics 按钮,展开 Statistics 对话框(见图 2-52)。在 Statistics 下框中选择 Harmonic Mean 将其移入 Cell Statistics 框中,要求计算调和均数。单击 Continue 按钮返回 Case Summaries 对话框。

④ 单击 OK 按钮执行,在输出窗口中出现计算结果,见表 2-35。

因此,服用一定剂量的缓叭妥后,大鼠的平均睡眠时间为 39.2 分钟。

例 2.37 水果甲级每元 1 公斤,乙级每元 1.5 公斤,丙级每元 2 公斤。问:

- 若各买 1 公斤,平均每元可买多少公斤?

表 2-35 调和均数

Case Summaries	
Harmonic Mean	
睡眠时间	
	39.2051

- 各买 6.5 公斤，平均每元可买多少公斤？
- 甲级 3 公斤，乙级 2 公斤，丙级 1 公斤，平均每元可买几公斤？

在 SPSS 中，解题步骤如下：

(1) 在 SPSS 中，建立数据文件，见 data02-32.sav

(2) 计算若各买 1 公斤，平均每元可买多少公斤

① 加权处理。按 Data→Weight Cases 顺序，打开 Weight Cases 对话框，见图 2-73。选择 Weight Cases by 选项，在左侧变量名源框中，选中购买公斤数 1 变量，并按右移箭头将购买公斤数 1 移入 Frequency Variable 框中，作为加权变量。单击 OK 按钮，完成加权设置。

② 按 Analyze→Reports→Case Summaries 顺序，打开 Case Summaries 对话框，见图 2-51。在左边的变量名源框中，选中 每元公斤数，将其移入 Variables 框中，关闭本对话框中的其他选项。

③ 单击 Statistics 按钮，展开 Statistics 对话框(见图 2-52)。在 Statistics 下框中选择 Harmonic Mean 将其移入 Cell Statistics 框中，要求计算调和均数。单击 Continue 按钮返回 Case Summaries 对话框。

④ 单击 OK 按钮执行，在输出窗口中出现计算结果，见表 2-36。

即各买 1 公斤时平均每元可买 1.3846 公斤。

(3) 计算若各买 6.5 公斤，平均每元可买多少公斤

除在上述(1)中将加权变量改为 购买公斤数2 外，其余步骤同上①至④，结果见表 2-37。各买 6.5 公斤，平均每元可买 1.3846 公斤。

(4) 计算买甲级 3 公斤、乙级 2 公斤和丙级 1 公斤时平均每元可买几公斤

除在上述(1)中将加权变量改为 购买公斤数3 外，其余步骤同上①至④，结果见表 2-38。买甲级 3 公斤、乙级 2 公斤和丙级 1 公斤时，平均每元可买 1.2414 公斤。

表 2-36 调和均数 (1)

Case Summaries	
	每元公斤数
1	1.00
2	1.50
3	2.00
Total Harmonic Mean	1.3846

表 2-37 调和均数 (2)

Case Summaries	
	每元公斤数
1	1.00
2	1.50
3	2.00
Total Harmonic Mean	1.3846

表 2-38 调和均数 (3)

Case Summaries	
	每元公斤数
1	1.00
2	1.50
3	2.00
Total Harmonic Mean	1.2414

6. 百分位数(Percentiles)

(1) 定义：百分位数是将一组有序数组（由小到大或由大到小排列的数据序列）分割为 100 等份，与 99 个分割点位置上相对应的变量值称为百分位数，分别记作 P_1, P_2, \dots, P_{99} 。

(2) 意义： P_1 表示有 1% 的数据落在其下， P_2 表示有 2% 的数据落在其下， \dots ， P_{99} 表示有 99% 的数据落在其下。

第 25 个百分数对应的观察值又称为第一四分位数，记作 Q_1 ，第 50 个百分位数为第二四分位数，记作 Q_2 ，第 75 个百分位数为第三四分位数，记作 Q_3 。

(3) 实例

例 2.38 求例 2.14 中实测的 100 名健康成年女子血清总蛋白含量的 2.5% 与 97.5% 对应的百分位数，原始数据已存放在 data02-18.sav 的数据文件中。

在 SPSS 中，解题步骤如下：

① 在 SPSS 中，打开数据文件 data02-18.sav。

② 按 Analyze→Descriptive Statistics→Frequencies 顺序，展开 Frequencies 对话框，见图 2-50。在变量名表中，选择所要分析的 *血清总蛋白* 变量，并将其移到 Variables 矩形框中。

③ 按 Statistics 按钮，弹出 Statistics 对话框，见图 2-93。

在 Percentile Values 下选择 Percentile(s) 选项，在其后框中，输入 2.5，单击 Add 按钮，将 2.5 百分位数增加到所要求计算的百分位数框中，再输入 97.5，单击 Add 按钮，将 97.5 百分位数增加到所要求计算的百分位数框中。

单击 Continue 按钮，返回 Frequencies 对话框。

取消系统默认选项 Display frequency tables。

④ 单击 OK 按钮运行，则在输出窗口中得到 2.5 和 97.5 的百分位数计算表，见表 2-39。

可见，健康成年女子血清总蛋白含量的 2.5% 与 97.5% 对应的百分位数分别为 6.50 和 8.16。

表 2-39 百分位数表

Statistics		
血清总蛋白		
N	Valid	100
	Missing	0
Percentiles	2.5	6.5000
	97.5	8.1600

2.3.3.2 变异性统计量

1. 两极差 (Range)

(1) 两极差又称全距。它是一组数据资料中最大值与最小值之差。

(2) 计算公式为： $R = x_{\max} - x_{\min}$

(3) 应用：它是离中趋势指标，它越大，反映数据资料越分散。其优点是计算方便，

只涉及原始数据中的两个值，但缺点是反映原始数据不够全面。

例 2.39 测得甲、乙两个玉米品种各 10 个的果穗长度 (cm) 如下，见表 2-40。试比较甲、乙两个玉米品种的变异性。

表 2-40 甲、乙两个玉米品种的果穗长度

甲	15	15	16	17	20	20	21	21	23	24
乙	17	17	18	19	19	20	20	20	21	21

在 SPSS 中，解题步骤如下：

① 在 SPSS 中，建立数据文件，见 data02-33.sav。

② 按 Analyze→Reports→Case Summaries 顺序，打开 Case Summaries 对话框，见图 2-51。在左边的变量名源框中，选中 **果穗长度**，将其移入 Variables 框中，选择 **品种** 变量，并将其移入 Grouping Variable[s] 框中，关闭本对话框中的其他选项。

③ 单击 Statistics 按钮，展开 Statistics 对话框（图 2-52）。在 Statistics 下框中选择 Minimum、Maximum、Range 将其移入 Cell Statistics 框中，要求计算最小值、最大值和两极差。单击 Continue 按钮返回 Case Summaries 对话框。

④ 单击 OK 按钮执行，在输出窗口中出现计算结果，见表 2-41。

由于甲组的两极差为 9 大于乙组的两极差 4，故，甲组的数据资料比乙组的数据资料变异性大。

2. 四分位间距 (Interquartiles Range)

(1) 定义：四分位间距又称为四分位差。它是将一组由小到大排列的有序变量值四等分后，用第三四分位数减去第一四分位数所得的差值，记为 $Q = Q_3 - Q_1$ 。

(2) 意义：它越小说明中间 50% 的数据越集中。其优缺点同两极差。

(3) 实例

例 2.40 求例 2.14 中实测的 100 名健康成年女子血清总蛋白含量的四分位间距，原始数据已存放在 data02-18.sav 的数据文件中。

在 SPSS 中，解题步骤如下：

① 在 SPSS 中，打开数据文件 data02-18.sav。

② 按 Analyze→Descriptive Statistics→Frequencies 顺序，展开 Frequencies 对话框，见图 2-50。在变量名表中，选择所要分析的 **血清总蛋白** 变量，并将其移到 Variables 矩形框中。

③ 单击 Statistics 按钮，弹出 Statistics 对话框，见图 2-93。

表 2-41 两极差

Case Summaries

果穗长度			
品种	Minimum	Maximum	Range
甲	15.00	24.00	9.00
乙	17.00	21.00	4.00
Total	15.00	24.00	9.00

在 Percentile Values 下选择 Quartiles 选项。

单击 Continue 按钮, 返回 Frequencies 对话框。

取消系统默认选项 Display frequency tables。

④ 单击 OK 按钮运行, 则在输出窗口中得到四分位数计算表, 见表 2-42。

由于第三四分位数为 7.58, 而第一四分位数为 7.12, 故健康成年女子血清总蛋白含量的四分位数间距为 $7.58 - 7.12 = 0.46$ 。

表 2-42 四分位数计算表

Statistics		
血清总蛋白		
N	Valid	100
	Missing	0
	Percentiles	25
		7.1200
	50	7.3500
	75	7.5800

3. 方差 (Variance) 和标准差 (Standard Deviation)

(1) 定义: 方差是所有变量值与其平均数偏差的平方和的平均值。而方差的算术平方根, 称标准差。

(2) 计算公式:

$$\text{总体方差: } \sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}, \quad \text{总体标准差: } \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$$

$$\text{样本方差: } S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}, \quad \text{样本标准差: } S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

若样本观察值 x_1, x_2, \dots, x_n , 分别出现了 f_1, f_2, \dots, f_n 次, 则定义 $s^2 = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\sum_{i=1}^n f_i - 1}$ 为样

本加权资料的方差, $s = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\sum_{i=1}^n f_i - 1}}$ 为样本加权资料的标准差。

(3) 应用: 它们是离中趋势指标, 方差越大, 则标准差也越大, 说明该组数据距离平均数这个中心的离散趋势越大。

(4) 实例

例 2.41 求例 2.14 中实测的 100 名健康成年女子血清总蛋白含量的方差和标准差, 原始数据已存放在 data02-18.sav 的数据文件中。

在 SPSS 中, 解题步骤如下:

① 在 SPSS 中, 打开数据文件 data02-18.sav。

② 按 Analyze→Descriptive Statistics→Frequencies 顺序, 展开 Frequencies 对话框,

见图 2-50。在变量名表中，选择所要分析的 *血清总蛋白* 变量，并将其移到 Variables 矩形框中。

③ 单击 Statistics 按钮，弹出 Statistics 对话框，见图 2-93。

在 Dispersion 下选择 Std. Deviation 和 Variance 选项。要求计算标准差和方差。

单击 Continue 按钮，返回 Frequencies 对话框。

取消系统默认选项 Display frequency tables。

表 2-43 方差和标准差表

④ 单击 OK 按钮运行，则在输出窗口中得到方差和标准差的计算表，见表 2-43。

从表中可见，健康成年女子血清总蛋白含量的标准差和方差分别为 0.39528 和 0.156。

4. 标准误 (S.E.mean)

(1) 定义：在用样本均数估计总体均数时，由于抽样误差的存在，各个样本均数间并不是完全相同的，由样本均数组成的新样本所求得的标准差，称为样本标准误。

(2) 计算公式为： $S_e = S / \sqrt{n}$

(3) 应用：它反映样本均数之间的离散程度。标准误越大，说明样本均值之间的越离散。

(4) 实例

例 2.42 求例 2.39 中测得甲、乙两个玉米品种各 10 个的果穗长度 (cm) 的标准误，原始数据存放在 data02-33.sav 的数据文件中。

在 SPSS 中，解题步骤如下：

① 在 SPSS 中，打开数据文 data02-33.sav。

② Analyze→Reports→Case Summaries 顺序，打开 Case Summaries 对话框，见图 2-51。在左边的变量名源框中，选中 *果穗长度*，将其移入 Variables 框中，选择 *品种* 变量，并将其移入 Grouping Variable[s] 框中，关闭本对话框中的其他选项。

表 2-44 标准误

Case Summaries	
Std. Error of Mean	
品种	果穗长度
甲	1.03064
乙	.46667
Total	.55060

③ 单击 Statistics 按钮，展开 Statistics 对话框（见图 2-52）。在 Statistics 下框中选择 Std. Error of Mean 将其移入 Cell Statistics 框中，要求计算标准误。单击 Continue 按钮返回 Case Summaries 对话框。

④ 单击 OK 按钮执行，在输出窗口中出现计算结果，见表 2-44。

由于甲组的标准误为 1.03064 大于乙组的标准误 0.46667，因此，用甲组的均值估计总体均值时要比用乙组的均值估计总体均值时的变异性大。

5. 变异系数

(1) 定义：变异系数是标准差与均数的百分比。用 C_v 表示。

(2) 计算公式：
$$C_v = \frac{s}{\bar{x}} \times 100\%$$

(3) 意义：变异系数是反映数据资料离散程度的指标，变异系数越大数据资料越离散。在生物学中，当变异系数小于 10% 或最多不大于 15% 时，它还可用来表示生物的同类型。

(4) 应用：单位量纲不同的数据资料离散程度的比较时，可用变异系数。

(5) 实例

例 2.43 以例 2.11 为例，试比较身高与足长的离中趋势。数据资料存放在 data02-16.sav 中。

在 SPSS 中，解题步骤如下：

① 在 SPSS 中，打开数据文件 data02-16.sav。

② Analyze→Descriptive Statistics→Ratio 顺序，打开 Ratio 对话框，见图 2-97。

在左侧变量名框中选择 **测试值**，将其移入 Numerator 框中，选择 **系数**，将其移入 Denominator 框中，即将 **测试值** 作为分子，将 **系数** 作为分母，选择 **变量类别**，将其移入 Group Variable 框中。

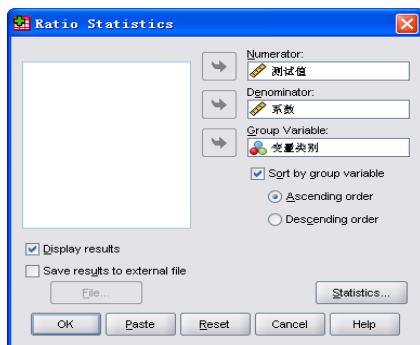


图 2-97 Ratio 对话框

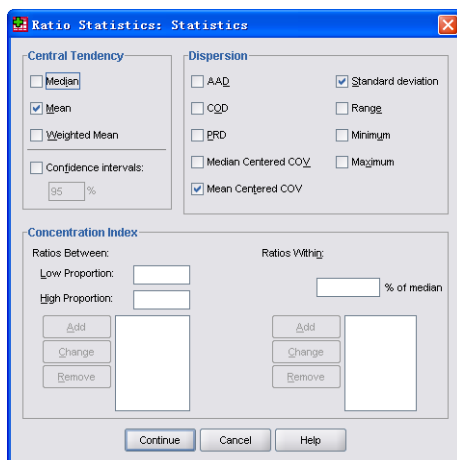


图 2-98 Statistics 对话框

③ 单击 Statistics 按钮，打开 Statistics 对话框，见图 2-98。

选择 Mean、Standard deviation、Mean Centered COV 选项，要求计算平均数、标准差和变异系数三个统计量。

单击 Continue 按钮，返回 Ratio 对话框。

④ 单击 OK 按钮执行，在输出窗口中，出现计算结果，见表 2-45。

表 2-45 变异系数结果

Ratio Statistics for 测试值/系数			
Group	Mean	Std. Deviation	Coefficient of Variation
			Mean Centered
身高	157.500	5.359	3.4%
足长	22.580	1.443	6.4%
Overall	90.040	69.318	77.0%

6. 结果与讨论

从表 2-45 可见, 身高的平均数为 157.5, 标准差为 5.359, 变异系数为 3.4%, 足长的平均数为 22.580, 标准差为 1.443, 变异系数为 6.4%, 由于足长的变异系数 6.4% > 身高的变异系数 3.4%, 说明 13 岁男孩的足长数据资料的离散程度比身高大。

2.3.4 分布形态

1. 偏度 (Skewness)

(1) 定义: 它是描述某变量取值分布对称性的统计量。是描述数据资料的分布形态的。

$$(2) \text{ 计算公式: } Skewness = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)S^3}$$

(3) 应用: 它是与正态分布相比较的量, 当它等于 0 时, 表示数据分布的形态与正态分布偏度相同; 当它大于 0 时, 表示正偏差数值较大, 为正偏, 表示右边拖着一条长尾巴; 当它小于 0 时, 表示负偏差数值较大, 为负偏, 表示左边拖着一条长尾巴。

(4) 实例

例 2.44 求例 2.14 中实测的 100 名健康成年女子血清总蛋白含量的偏度, 原始数据已存放在 data02-18.sav 的数据文件中。

在 SPSS 中, 解题步骤如下:

① 在 SPSS 中, 打开数据文件 data02-18.sav。

② 按 Analyze → Descriptive Statistics → Frequencies 顺序, 展开 Frequencies 对话框, 见图 2-50。在变量名表中, 选择所要分析的 血清总蛋白 变量, 并将其移到 Variables 矩形框中。

③ 单击 Statistics 按钮, 弹出 Statistics 对话框, 见图 2-93。

在 Distribution 下选择 Skewness 选项。要求计算偏度。

单击 Continue 按钮，返回 Frequencies 对话框。

取消系统默认选项 Display frequency tables。

④ 单击 OK 按钮运行，则在输出窗口中得到偏度的计算表，见表 2-46。

从表中可见，健康成年女子血清总蛋白含量的偏度为 0.077，偏度的标准误为 0.241。由于偏度接近于 0，故分布形态与正态分布基本一致。

2. 峰度 (Kurtosis)

(1) 定义：描述某变量所有取值分布形态陡缓程度的统计量。

$$(2) \text{ 计算公式: } Kurtosis = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{(n-1)S^4} - 3$$

(3) 应用：它是与正态分布相比较的量，它为 0 时，表示其数据分布与正态分布的陡缓程度相同；它大于 0 时，表示其数据分布比正态分布的峰度要更加陡峭，为尖顶峰；它小于 0 时，表示其数据分布比正态分布的峰度要更加平坦，为平顶峰。

(4) 实例

例 2.45 求例 2.14 中实测的 100 名健康成年女子血清总蛋白含量的峰度，原始数据已存放在 data02-18.sav 的数据文件中。

在 SPSS 中，解题步骤如下：

① 在 SPSS 中，打开数据文件 data02-18.sav。

② 按 Analyze→Descriptive Statistics→Frequencies 顺序，展开 Frequencies 对话框，见图 2-50。在变量名表中，选择所要分析的血清总蛋白变量，并将其移到 Variables 矩形框中。

③ 单击 Statistics 按钮，弹出 Statistics 对话框，见图 2-93。

在 Distribution 下选择 Kurtosis 选项。要求计算峰度。

单击 Continue 按钮，返回 Frequencies 对话框。

取消系统默认选项 Display frequency tables。

④ 单击 OK 按钮运行，则在输出窗口中得到峰度的计算表，见表 2-47。

从表中可见，健康成年女子血清总蛋白含量的峰度为 0.052，峰度的标准误差为 0.478。由于峰度接近于 0，故分布形态的高峰与正态分布基本一致。

表 2-46 偏度计算表

Statistics		
血清总蛋白		
N	Valid	100
	Missing	0
Skewness		.077
Std.Error of Skewness		.241

表 2-47 峰度计算表

Statistics		
血清总蛋白		
N	Valid	100
	Missing	0
Kurtosis		.052
Std.Error of kurtosis		.478

2.4 探索分析

2.4.1 探究分析的意义

探究分析是一切数据分析的前提和基础。通过探索分析可以了解原始数据中是否有异常数据, 是否有缺失值, 是否有可疑编码, 是否有明显的错误, 是否服从正态分布, 变量之间是否有相关性, 方差是否齐性等, 这样才能正确引导研究者使用正确的描述和推断统计的方法, 能够避免出现错误结论。此外通过对变量计算描述统计量和绘制统计图, 使研究人员对数据的概貌有一个较为全面的了解。

2.4.2 实例分析

1. 打印数据文件中的数据资料并检查缺失数据

例 2.46 跨栏技术、栏感、栏间节奏、第一个栏、后三个栏、分配体力、心理素质等 7 个因素, 一般被认为是影响跨栏成绩的重要因素。为研究这 7 个因素对跨栏成绩的重要性, 调查人员对 20 位资深的跨栏教练通过问卷调查的方式, 让他们对这 7 个因素进行排序, 7 代表最重要, 6 次之, 其余类推。现得到整理后的调查结果的数据资料, 见表 2-48。现已将该数据资料在 SPSS 中制成数据文件, 见 data02-34.sav。请打印数据文件中的数据资料, 并进行缺失数据检查。

在 SPSS 中, 能满足本例要求的 SPSS 过程远不止一个, 本例利用 Reports 中的 Case Summaries 来完成。具体步骤如下:

表 2-48 对影响跨栏因素的排序调查结果

编号	栏技术	栏感	栏间节奏	第一个栏	后三个栏	分配体力	心理素质
1	6	5	3	4	2	1	7
2	5	4	2	1	3	6	7
3	1	4	2	5	6	3	7
4	6	4	3	2	1	5	7
5	4	5	2	7	6	3	1
6	3	4	1	6	7	2	5
7	3	7	2	1	4	5	6
8	3	1	2	5	4	7	6
9	1		2			3	4
10	3	4	2	5	1	6	7
11		1	2	3	4	5	6

(续表)

编号	栏技术	栏感	栏间节奏	第一个栏	后三个栏	分配体力	心理素质
12			1	2	3	4	
13	3	4	2	6	7	1	8
14	5	3	4	6	1	2	7
15			1		2		
16			1				
17		2	3				3
18	2	4	3	5	6	1	7
19	1	2	3	5	4	1	7
20	2	4	3	5	6	1	7

① 打开 data02-34.sav。

② 按 Analyze→Reports→Case Summaries 顺序, 打开 Case Summaries 对话框, 见图 2-51。在左边的变量名源框中, 选中除编号以外的其他所有变量, 通过中间的右移箭头将它们移到 Variables 框中, 用同样的方法将编号变量移入 Grouping Variable(s)框中。在对话框的选择中 Display cases, 单击 Statistics 按钮, 打开 Statistics 对话框见图 2-52。

注意: 编号变量必须是字符型变量。

③ 在 Statistics 选项卡中将 Cell Statistics 下框中的统计量全部移回 Statistics 下框中。单击 Continue 按钮返回对话框。

④ 在对话框中, 单击 OK 按钮, 在输出窗口中得到运行结果, 见表 2-49, 表 2-50。

表 2-49 样品处理汇总

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
跨栏技术*编号	15	75.0%	5	25.0%	20	100.0%
栏感*编号	16	80.0%	4	20.0%	20	100.0%
栏间节奏*编号	20	100.0%	0	.0%	20	100.0%
第一个栏*编号	16	80.0%	4	20.0%	20	100.0%
后三个栏*编号	17	85.0%	3	15.0%	20	100.0%
分配体力*编号	17	85.0%	3	15.0%	20	100.0%
心理素质*编号	17	85.0%	3	15.0%	20	100.0%

表 2-50 数据文件中全部数据

Case Summaries									
			跨栏技术	栏感	栏间节奏	第一个栏	后三个栏	分配体力	心理素质
编号	1.00	1	6.00	5.00	3.00	4.00	2.00	1.00	7.00
	2.00	1	5.00	4.00	2.00	1.00	3.00	6.00	7.00
	3.00	1	1.00	4.00	2.00	5.00	6.00	3.00	7.00
	4.00	1	6.00	4.00	3.00	2.00	1.00	5.00	7.00
	5.00	1	4.00	5.00	2.00	7.00	6.00	3.00	1.00
	6.00	1	3.00	4.00	1.00	6.00	7.00	2.00	5.00
	7.00	1	3.00	7.00	2.00	1.00	4.00	5.00	6.00
	8.00	1	3.00	1.00	2.00	5.00	4.00	7.00	6.00
	9.00	1	1.00		2.00			3.00	4.00
	10.00	1	3.00	4.00	2.00	5.00	1.00	6.00	7.00
	11.00	1		1.00	2.00	3.00	4.00	5.00	6.00
	12.00	1			1.00	2.00	3.00	4.00	
	13.00	1	3.00	4.00	2.00	6.00	7.00	1.00	8.00
	14.00	1	5.00	3.00	4.00	6.00	1.00	2.00	7.00
	15.00	1			1.00		2.00		
	16.00	1			1.00				
	17.00	1		2.00	3.00				3.00
	18.00	1	2.00	4.00	3.00	5.00	6.00	1.00	7.00
	19.00	1	1.00	2.00	3.00	5.00	4.00	1.00	7.00
	20.00	1	2.00	4.00	3.00	5.00	6.00	1.00	7.00

⑤ 结果与讨论

表 2-49 从左向右依次列出了变量名、样品中有效观测的数量、它占总样品数（样本含量）的百分比、样品中无效观测的数量、它占总样品数的百分比、有效观测和无效观测的数量总和及它同总的样品数的百分比。

从样品（Cases）的无效观测（Excluded）列的 N 中，可以看到，除跨栏节奏变量上没有缺失之外，其他 6 个变量上都有缺失值，依次为跨栏技术中有 5 个缺失值，栏感和第一个栏中有 4 个缺失值，而后三个栏、分配体力、心理素质中各有 3 个缺失值。

表 2-50 为录入数据文件中的数据，它与原始记录数据间没有差异。

由此可见，数据文件中的缺失值，不是录入数据过程中造成的，而是被试对象没有填写的结果。

2. 检查异常数据

例 2.47 番茄汁罐头生产标准规定,100g 的罐头中,维生素 C 的含量不得少于 21mg,从某番茄汁罐头生产厂生产的一批罐头中随机抽取 22 个,测得维生素 C 的含量(单位:mg) 如下:

16, 22, 21, 20, 23, 21, 19, 15, 13, 23, 17, 20, 29, 18, 22, 16, 25, 12, 32, 10, 35, 36

试对该批数据作异常值检查。

在 SPSS 中进行异常值检查的步骤如下:

- ① 在 SPSS 数据编辑窗口中, 打开上述原始数据所做成的数据文件 data02-35.sav。
- ② 按 Analyze→Descriptive Statistics→Explore 顺序, 打开 Explore 对话框, 见图 2-99。在左边的变量名源框中, 选中维生素 C 变量, 通过中间的右移箭头将它们移到 Dependent List 框中, 单击 Statistics 按钮, 打开 Statistics 选项卡, 见图 2-100。

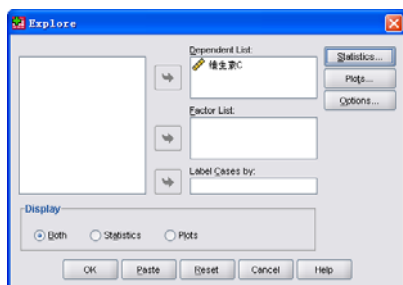


图 2-99 Explore 对话框

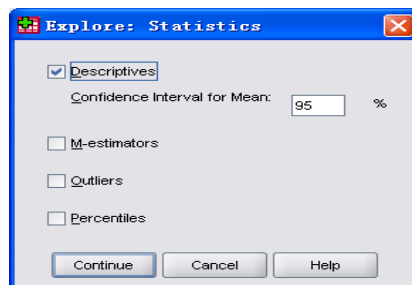


图 2-100 Statistics 选项卡

③ 在 Statistics 选项卡中, 选择 Descriptives 选项, 单击 Continue 按钮返回 Explore 对话框。其他保持系统默认选择。

④ 单击 OK 按钮运行, 在输出窗口中得到计算结果, 见表 2-51、图 2-101、图 2-102。

维生素C Stem-and-Leaf Plot

Frequency	Stem & Leaf
3.00	1 . 023
6.00	1 . 566789
8.00	2 . 00112233
2.00	2 . 59
1.00	3 . 2
2.00	Extremes (>=35)
Stem width: 10.00	
Each leaf: 1 case(s)	

图 2-101 茎、叶图

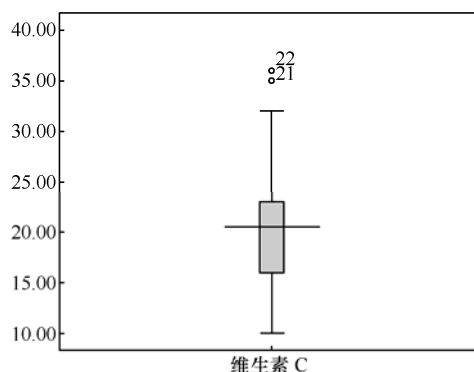


图 2-102 箱图

表 2-51 描述统计

Descriptives			Statistic	Std. Error
维生素 C	Mean		21.1364	1.48192
	95% Confidence Interval for Mean	Lower Bound	18.0545	
		Upper Bound	24.2182	
	5% Trimmed Mean		20.9242	
	Median		20.5000	
	Variance		48.314	
	Std. Deviation		6.95082	
	Minimum		10.00	
	Maximum		36.00	
	Range		26.00	
	Interquartile Range		7.50	
	Skewness		.700	.491
	Kurtosis		.150	.953

⑤ 结果与讨论

在表 2-51 中, 从上到下依次列出了变量维生素 C 的平均数、标准误分别为 21.1364、1.48192, 95%的置信区间的下限为 18.0545、上限为 24.2182, 5%的截尾(调整)平均数为 20.9242, 中位数为 20.5000, 方差为 48.314, 标准差为 6.95082, 最小值为 10, 最大值为 36, 两极差为 26, 四分位数间距为 7.50, 偏度及其标准误分别为 0.700、0.491, 峰度及其标准误分别为 0.150、0.953。

图 2-101 显示了原始数据资料的茎、叶图, 茎宽(Stem width)为 10, 每叶(Each leaf)代表一个样品, 频数(Frequency)给出的是一个区间中, 观察值出现的样品数, 所以, 茎(Stem)显示的是十分位上的数, 叶(Leaf)显示的是个位上的数, 由此可见, 10-13 之间, 一共出现了 3 个数, 分别是 10、12、13。其余类推, 极端值(大于等于 35)有 2 个。

图 2-102 显示了维生素 C 数据资料正常值的范围和平均数所在的位置, 同茎叶图反映的情况一样, 有两个异常值落在了正常范围之外, 分别是对应第 21 和第 22 个样品所在的观察值。

⑥ 结论

在维生素 C 变量上, 有两个异常值, 第 21 位的 35 和第 22 位的 36, 需要从专业的角度来判定其是否合理。如果不合理, 则应将它当作异常值予以剔除。

3. 检查计量资料的正态性和方差齐性

数据资料正态性的检验有多种方法,各种方法有不同的条件要求,详见第3章例3.13中的说明。

例 2.48 为研究不同的抗原对恶性疟原虫阳性患者间接荧光抗体的测试效果,研究人员将全部患者随机分配到3种不同的抗原组中,测定结果如下(见表2-52):

表 2-52 不同抗原体的测定结果

恶性疟原虫抗原组	1:160	1:160	1:320	1:320	1:640	1:640	1:640	1:640	1:1280	1:1280
诺氏疟原虫抗原组	1:80	1:160	1:160	1:160	1:160	1:320	1:320	1:320	1:640	1:640
食蟹猴疟原虫抗原组	1:40	1:80	1:80	1:80	1:160	1:160	1:160	1:160	1:320	1:320

试对以上数据资料做正态性检验。

【题析】由于测试值都是比例数据,所以在 SPSS 中建立数据文件时,只录入比例后面的数据,建成的数据文件见 data02-36.sav。其数据结构见图 2-103。

在 SPSS 中的操作步骤如下:

- ① 在 SPSS 数据编辑窗口打开数据文件 data02-36.sav。
- ② 按 Analyze→Descriptive Statistics→Explore 顺序,打开 Explore 对话框(见图 2-99)。
- ③ 在左边的变量名源框中,选中抗体水平变量,通过中间的右移箭头将它们移到 Dependent List 框中,用同样的方法将变量抗体类型,移到 Factor List 框中。在 Display 选项中,选择 Plots,按 Plots 按钮,打开 Plots 对话框,见图 2-104。

	抗原类型	抗体水平
1	1	160.00
2	1	160.00
3	1	320.00
4	1	320.00
5	1	640.00
6	1	640.00
7	1	640.00
8	1	640.00
9	1	1280.00
10	1	1280.00
11	2	80.00
12	2	160.00
13	2	160.00
14	2	160.00
15	2	160.00

图 2-103 录入的原始数据

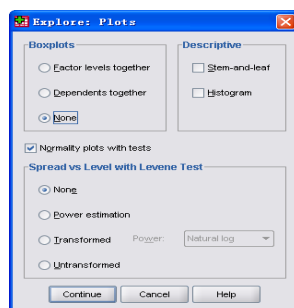


图 2-104 Plots 对话框

④ 在 Plots 选项卡中,在 Boxplots 选择项中,选择 None,不输出箱图。选择 Normality Plots with tests 选项,做正态分布检验。在 Descriptive 中关闭所有选项。其他保持系统默认选项。单击 Continue 按钮,返回 Explore 对话框。

⑤ 单击 OK 按钮运行,在输出窗口中得到计算结果,见表 2-53、表 2-54、图 2-105、图 2-106、图 2-107、图 2-108、图 2-109、图 2-110。

表 2-53 样品处理汇总

		Case Processing Summary					
		Cases					
		Valid		Missing		Total	
抗原类型		N	Percent	N	Percent	N	Percent
抗体水平	恶性疟原虫抗原组	10	100.0%	0	.0%	10	100.0%
	诺氏疟原虫抗原组	10	100.0%	0	.0%	10	100.0%
	食蟹猴疟原虫抗原组	10	100.0%	0	.0%	10	100.0%

表 2-54 正态分布检验

		Tests of Normality					
		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
抗原类型		Statistic	df	Sig.	Statistic	df	Sig.
抗体水平	恶性疟原虫抗原组	.268	10	.040	.849	10	.057
	诺氏疟原虫抗原组	.263	10	.048	.816	10	.023
	食蟹猴疟原虫抗原组	.284	10	.022	.840	10	.044

a. Lilliefors Significance Correction

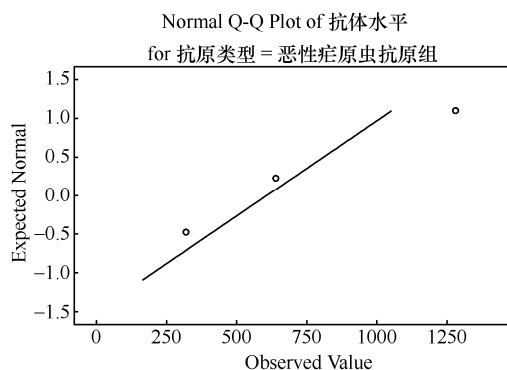


图 2-105 恶性疟原虫抗原组正态 Q-Q 图

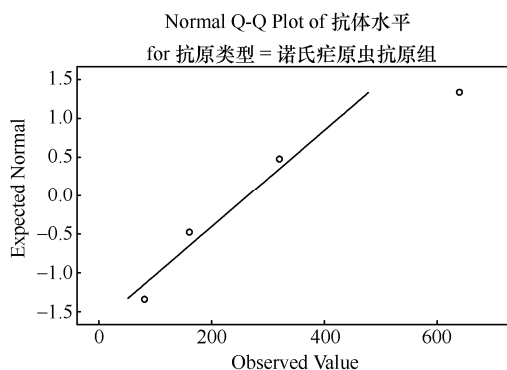


图 2-106 诺氏疟原虫抗原组正态 Q-Q 图

⑥ 结果与讨论

表 2-53 显示三个抗体水平的有效观察数据为 10，即三个样本的有效含量均为 10。

表 2-54 显示三个抗体水平正态分布检验的结果。Kolmogorov-Smirnov 统计量的值分别为：0.268、0.263、0.284，Lilliefors 显著性水平（Sig.）分别为：0.040、0.048、0.022。表明三个水平对应的样本数据服从正态分布的原假设成立的概率都小于 0.05，因此，拒绝原假设，而认为它们都不服从正态分布，此时，拒绝原假设犯错误的概率不到 0.05。

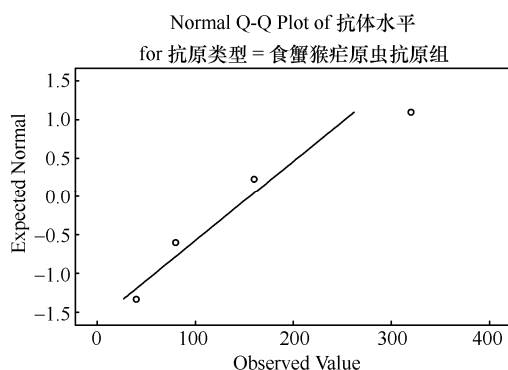


图 2-107 食蟹猴疟原虫抗原组正态 Q-Q 图

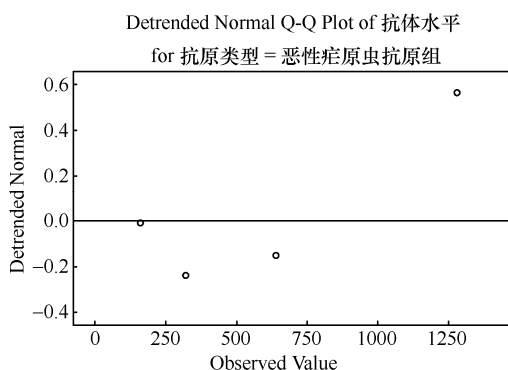


图 2-108 恶性疟原虫抗原组正态 Q-Q 偏差图

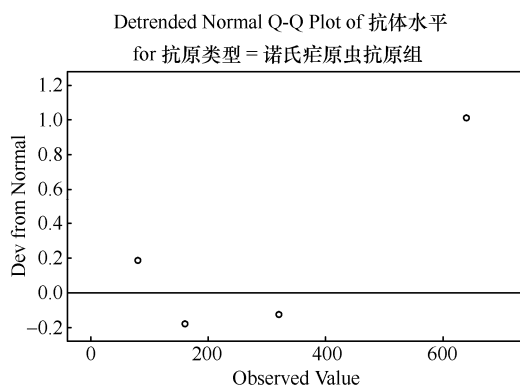


图 2-109 诺氏疟原虫抗原组正态 Q-Q 偏差图

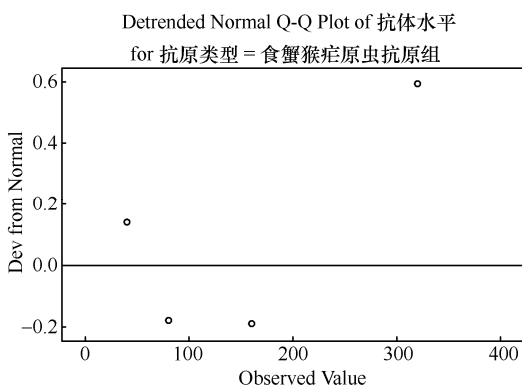


图 2-110 食蟹猴疟原虫抗原组正态 Q-Q 偏差图

从正态 Q-Q 图上也可看出每个水平上都有数据明显偏离直线，在 Q-Q 偏差图上，偏差数据明显不对称，这从另一个方面再次说明原始数据资料不服从正态分布。

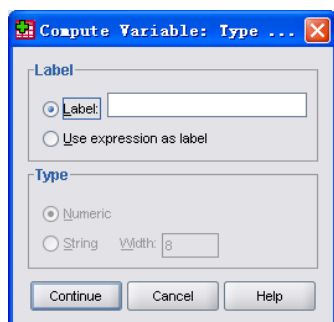


图 2-111 定义标签和类型

⑦ 由上可知，三个水平对应样本的原始数据不服从正态分布，根据原始数据资料有成等比数列的趋势，因此，可用以 10 为底的对数将其转换，再对转换后的数据资料进行正态分布检验。故在 SPSS 中，继续如下操作步骤进行对数正态分布检验。

⑧ 按 Transform→Compute Variable 顺序，打开 Compute Variable 对话框，见图 2-53。在 Target Variable 框中，输入目标变量名为 抗体水平转换值，按 Type & Label 按钮，打开 Type & Label 选项卡，见图 2-111。

选择 Use expression as label，用表达式中变量名作

为标签。在 Type 中,用系统默认选项 Numeric。单击 Continue 按钮,返回 Compute Variable 对话框。

⑨ 在 Function groups 框中,选择 All,在 Function and Special Variables 框中,选中 Lg10 函数,双击鼠标左键,则在 Numeric Expression 框中出现“LG10(?)”,在左边变量名源框中选中 抗体水平变量,按右移箭头,将 抗体水平 代替 LG10(?)中的“?”,见图 2-112。

⑩ 单击 OK 按钮运行,则在数据编辑窗口中,出现一系列变量名为 抗体水平转换值的转换值。见图 2-113。

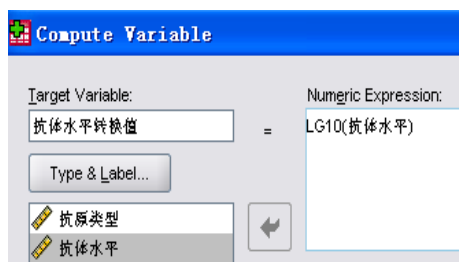


图 2-112 表达式的设定

	抗原类型	抗体水平	抗体水平转换值
1	1	160.00	2.20
2	1	160.00	2.20
3	1	320.00	2.51
4	1	320.00	2.51
5	1	640.00	2.81
6	1	640.00	2.81
7	1	640.00	2.81
8	1	640.00	2.81
9	1	1280.00	3.11
10	1	1280.00	3.11
11	2	80.00	1.90

图 2-113 数值对数转换结果

⑪ 按 Analyze→Descriptive Statistics→Explore 顺序,打开 Explore 对话框,见图 2-99。

⑫ 在左边的变量名源框中,选中 抗体水平转换值变量,通过中间的右移箭头将它们移到 Dependent List 框中,用同样的方法将变量 抗原类型,移到 Factor List 框中。在 Display 选项中,选择 Plots,单击 Plots 按钮,打开 Plots 对话框,见图 2-104。

⑬ 在 Plots 选项卡中,在 Boxplots 选择项中,选择 None,不输出箱图。选择 Normality Plots with tests 选项,做正态分布检验。在 Descriptive 中关闭所有选项。

⑭ 选择 Spread vs Level with Levene Test,做方差的 Levene 齐性检验,输出散布-层次图,包括回归的幂和回归直线的斜率。当没有指定因素的分组变量时,此选项无效。

关于 Levene 齐性检验的基本原理参见 6.1.2。

它有四个选项:

- None: 不做齐性检验,也不输出散布-层次图。
- Power estimation: 转换幂值估计,表示对每一组数据产生一个中位数范围的自然对数与四分位范围的自然对数的散点图。
- Transformed: 有使用者在“Power”下拉框中选择幂变换使用的幂值,对原始数据进行转换。可以选择的幂值有:自然对数(Natural log)、平方根分之一(1/Square root)、倒数(Reciprocal)、平方根(Square root)、平方(Square)、立方(Cube)。

• Untransform: 对原始数据不做转换。

本例中由于已对原始数据做过对数转换,所以选择其中的 Untransform 选项。

单击 Continue 按钮，返回 Explore 对话框。

⑮ 单击 OK 按钮运行，在输出窗口中得到计算结果，在输出窗口中计算结果，见表 2-55、表 2-56、表 2-57、图 2-114、图 2-115、图 2-116、图 2-117、图 2-118、图 2-119、图 2-120。

表 2-55 样品处理汇总

Case Processing Summary							
抗原类型		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
COMPUTE 抗体水平转	恶性疟原虫抗原组	10	100.0%	0	.0%	10	100.0%
换值=LG10（抗体水平）	诺氏疟原虫抗原组	10	100.0%	0	.0%	10	100.0%
	食蟹猴疟原虫抗原组	10	100.0%	0	.0%	10	100.0%

表 2-56 正态分布检验

Tests of Normality							
抗原类型		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
COMPUTE 抗体水平转	恶性疟原虫抗原组	.245	10	.090	.892	10	.177
换值=LG10（抗体水平）	诺氏疟原虫抗原组	.222	10	.178	.906	10	.258
	食蟹猴疟原虫抗原组	.224	10	.168	.911	10	.287

Liuliefors Significance Correction

表 2-57 方差齐性检验

Tests of Homogeneity of Variance					
		Levene Statistic	Df1	Df2	Sig.
COMPUTE 抗体水平转	Based on Mean	.345	2	27	.712
换值=LG10（抗体水平）	Based on Median	.080	2	27	.924
	Based on Median and with adjusted df	.080	2	22.653	.924
	Based on trimmed mean	.334	2	27	.719

⑯ 结果与讨论

从表 2-56 可见，转换后的值由于服从正态分布的原假设成立的概率都大于 0.05，所以没有足够的证据推翻原假设，即认为数据资料服从对数正态分布。

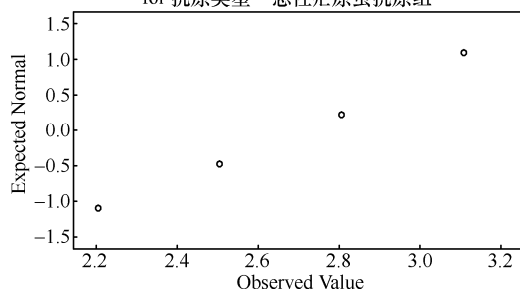
Normal Q-Q Plot of COMPUTE 抗体水平转换值=LG10(抗体水平)
for 抗原类型 = 恶性疟原虫抗原组

图 2-114 恶性疟原虫抗原组转换值正态 Q-Q 图

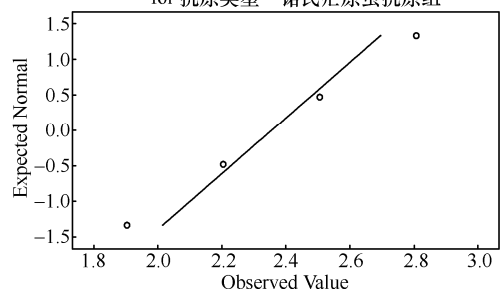
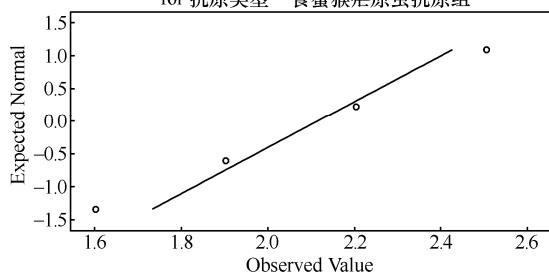
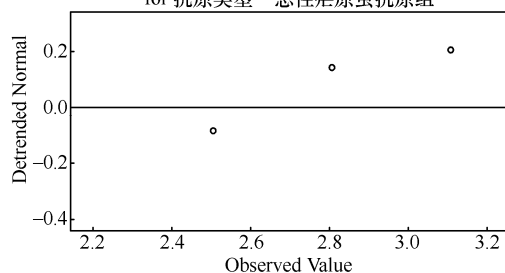
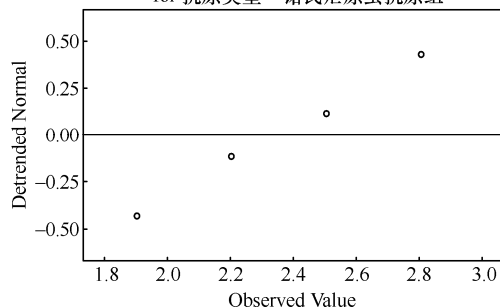
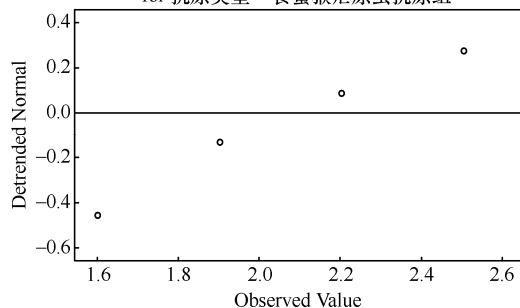
Normal Q-Q Plot of COMPUTE 抗体水平转换值=LG10(抗体水平)
for 抗原类型 = 诺氏疟原虫抗原组

图 2-115 诺氏疟原虫抗原组转换值正态 Q-Q 图

Normal Q-Q Plot of COMPUTE 抗体水平转换值=LG10(抗体水平)
for 抗原类型 = 食蟹猴疟原虫抗原组图 2-116 食蟹猴疟原虫抗原组
转换值正态 Q-Q 图Normal Q-Q Plot of COMPUTE 抗体水平转换值=LG10(抗体水平)
for 抗原类型 = 恶性疟原虫抗原组图 2-117 恶性疟原虫抗原组
转换值正态 Q-Q 偏差图Detrended Normal Q-Q Plot of COMPUTE 抗体水平转换值=LG10(抗体水平)
for 抗原类型 = 诺氏疟原虫抗原组图 2-118 诺氏疟原虫抗原组
转换值正态 Q-Q 偏差图Detrended Normal Q-Q Plot of COMPUTE 抗体水平转换值=LG10(抗体水平)
for 抗原类型 = 食蟹猴疟原虫抗原组图 2-119 食蟹猴疟原虫抗原组
转换值正态 Q-Q 偏差图

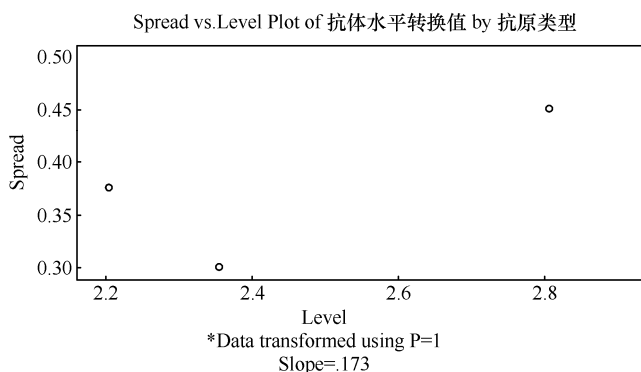


图 2-120 与抗原类型相对应的抗体水平转换值的分布图

从表 2-57 可见，方差齐性检验的结果表明，方差齐性的原假设成立的概率为 0.712 远远大于 0.05，所以没有理由拒绝方差齐性的假设。

图 2-114 至图 2-116 显示了转换值的正态 Q-Q 图，图中显示所有的散点基本都在一条直线上，图 2-117 至图 2-119 显示了转换值的正态 Q-Q 偏差图，偏差散点均匀分布在直线两侧，故显示对数转换值基本服从正态分布，与检验结果相一致。

图 2-120 显示，幂=1、斜率=0.173 时的与抗原类型相对应的抗体水平转换值的散布图。

2.5 计算派生指标

在天气预报、大学排名、经济景气指数、综合国力等预测、综合评估等研究分析中，经常要用到用已知变量的测试值去计算一个或多个新变量的值。统称这样的一个过程为计算派生指标。

例 2.49 德国学者劳雷尔于 1908 年提出，人体是个立方体，身高是立方体的一个边，用体重除以身高的立方，可反映骨骼、肌肉内脏器官及脂肪组织的发育情况，它是显示人体充实程度和营养状况的指数。他定义的计算公式如下：

$$\text{劳雷尔指数} = \text{体重} / \text{身高}^3 \times 100^2$$

其中：体重的单位为克，身高单位为厘米。

现从某校初一 13 岁男生中随机抽取 29 位男生，测得他们的身高、体重的资料，已存放在 data02-37.sav 中，试计算其劳雷尔指数。

在 SPSS 中，具体的操作步骤如下：

1. 在 SPSS 数据编辑窗口中打开 data02-37.sav 数据文件。
2. 按 Transform→Compute Variable 顺序打开 ComputeVariable 对话框，见图 2-53。

在 Target Variable 的下框中输入目标变量名 *劳雷尔指数*。

3. 在 Numeric Expression 的下框中输入：体重 / 身高 / 身高/身高*100*100。

它可以直接用键盘输入，也可以按下述操作过程来实现：

(1) 单击变量名源框中的体重，单击右移箭头按钮，将其移入 Numeric Expression 的下框中，单击对话框中的 “/” 按钮；

(2) 单击变量名源框中的身高，单击右移箭头按钮，将其移入 Numeric Expression 的下框中，单击对话框中的 “/” 按钮；

(3) 单击变量名源框中的身高，单击右移箭头按钮，将其移入 Numeric Expression 的下框中，单击对话框中的 “/” 按钮；

(4) 单击变量名源框中的身高，单击右移箭头按钮，将其移入 Numeric Expression 的下框中，单击 Shift+8 (输入乘号)，输入 100，再按 Shift+8 (输入乘号)，输入 100，完成计算新变量的表达式的输入。

4. 单击 OK 运行，则在当前工作的数据编辑窗口中，生成名为劳雷尔指数的新变量，并填充其计算结果，见图 2-121。

	编号	身高	体重	劳雷尔指数
1	1	135.10	32.0	0.13
2	2	139.90	30.4	0.11
3	3	163.60	46.2	0.11
4	4	146.50	33.5	0.11
5	5	156.20	37.1	0.10
6	6	156.40	35.5	0.09

图 2-121 计算派生指标

2.6 排 名

在央视和地方电视台举办的评定性的比赛以及体育的评分性项目的比赛中，都要用表演者的最后得分进行排名，在综合性评定中也要用到用最后得分进行排名。这可以使用 SPSS 中的 Transform 中的 Rank Cases 过程来实现。

例 2.50 为对 1993 年全运会女子争先赛决赛的运动员的人选进行预测，使训练更有针对性，科研人员在专家调查的基础上，得出了运动员在大型综合性运动会上取得优异成绩主要因素为近期主要比赛的成绩、保持高水平运动的年限和年龄因素，对这些因素换算成了相应的得分，并获得了各自的权重。得到的综合评估方程为：

$$Y = \text{青运会名次得分} + \text{八九年锦标赛名次得分} + 0.5 \times \text{八九年冠军赛名次得分} + 0.8 \times \text{八八年锦标赛名次得分} + \text{年龄因素得分} + 0.8 \times \text{已保持高水平运动年限得分}$$

运动员各影响因素得分已存放在数据文件 data02-38 中，试给各运动员进行排名。

在 SPSS 中计算的操作步骤如下:

(1) 在 SPSS 数据编辑窗口中打开 data02-38.sav 数据文件。

(2) 按 Transform→Compute Variabe 顺序打开 Compute Variabe 对话框, 见图 2-53。在 Target Variable 的下框中输入目标变量名综合得分。

(3) 在 Numeric Expression 的下框中直接输入: 八九年锦标赛得分+八九年冠军赛得分*0.5+第二届青运会决赛得分+八八年锦标赛得分*0.8+年龄因素得分+已保持高水平运动年限得分*0.8。

(4) 单击 OK 按钮, 在当前工作的数据编辑窗口中, 生成名为综合得分的新变量及其计算结果。

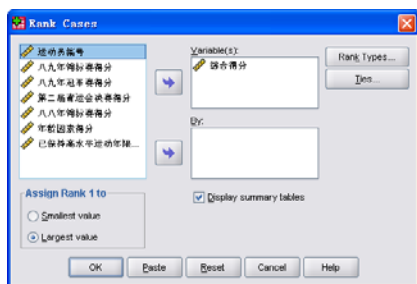


图 2-122 Rank Cases 对话框

(5) 按 Transform→Rank Cases 顺序, 打开 Rank Cases 对话框, 见图 2-122。在变量名列框中, 选择综合得分, 并用右移按钮将其移到 Variables 框中。

(6) 在 Assign Rank 1 to (将秩 1 赋予) 选项中, 选择 Largest Value (最大值)。

(7) 单击 OK 按钮, 在当前工作的数据文件中出现 R 综合得分排名。见图 2-123。

(8) 结论

R 综合得分排名前八名的运动员编号分别是:

1、2、9、6、3、7、8、4。

	运动员编号	八九年锦标赛得分	八九年冠军赛得分	第二届青运会决赛得分	八八年锦标赛得分	年龄因素得分	已保持高水平运动年限得分	综合得分	R综合得
1	1	6.00	6.00	0.00	8.00	7.00	3.00	24.80	1.000
2	2	8.00	8.00	0.00	7.00	3.00	1.00	21.40	2.000
3	3	7.00	3.00	0.00	0.00	8.00	5.00	20.50	5.000
4	4	5.00	0.00	0.00	0.00	8.00	5.00	17.00	8.000
5	5	4.00	7.00	0.00	3.00	6.00	1.00	16.70	9.000
6	6	3.00	4.00	0.00	6.00	8.00	4.00	21.00	4.000
7	7	0.00	0.00	8.00	0.00	8.00	5.00	20.00	6.000
8	8	0.00	0.00	7.00	0.00	8.00	5.00	19.00	7.000
9	9	0.00	0.00	6.00	5.00	8.00	4.00	21.20	3.000
10	10	0.00	0.00	5.00	0.00	7.00	5.00	16.00	10.500
11	11	0.00	0.00	4.00	0.00	8.00	5.00	16.00	10.500
12	12	0.00	0.00	3.00	0.00	8.00	5.00	15.00	12.000
13	13	0.00	0.00	2.00	0.00	8.00	5.00	14.00	14.000
14	14	0.00	0.00	1.00	0.00	8.00	5.00	13.00	15.000
15	15	0.00	0.00	0.00	4.00	7.00	5.00	14.20	13.000
16	16	0.00	5.00	0.00	0.00	3.00	0.00	5.50	16.000

图 2-123 综合得分计算结果

2.7 分析多重应答集

多重应答就是我们俗称的多选题，它是社会学、心理学、流行病学等问卷调查和市场研究中最常见的一种数据记录形式。例如，人们对颜色的喜好，有人喜欢黄、红，有人喜欢蓝、绿，甚至有人对所有颜色都喜欢，当我们列出所有颜色的种类，而让被试者选出其所有喜欢的颜色时，便得到了多重应答集。

在第1章1.2节中已经提到过，对于多选题，应建立等同于选项数的变量，每个变量只有两个可选结果，即是（编码1）、不是（编码0），故它也被称为多重二分法（multiple dichotomy method）。而对于排序题，它类似多选题，而又不同于多选题，因为它需要对所选结果进行排序，因此，当对所有选项都要进行排序时，就也要建立等同于选项数的变量，而此时变量的取值就是多选项之一，它就是多重分类法。

下面通过两个实例介绍在 SPSS 中，建立多重应答集并进行分析的过程。

2.7.1 多选题的处理

例 2.51 在对北京市 5 所体校运动员的问卷调查研究中，其中有一道题为“如果你能在大学中继续从事你的运动专项，在下面的选项中，你最想去的地方是：（可多选）”

（1）大学高水平运动队（2）体育系（3）运动系（4）军警校（5）职业学院（6）其他

数据文件存放在 data02-01.sav 中的 Q5.1-Q5.6 中。请建立多重应答集，并进行分析。

在 SPSS 中，建立多重应答集和进行多重应答集分析过程如下：

① 在 SPSS 数据编辑窗口中，打开 data02-01.sav。

② 按 Analyze→Tables→ Multiple Response Sets 顺序，或按 Analyze→Multiple Response→Define Variable Sets，打开 Define Multiple Response Sets 对话框，见图 2-124。

这两种方法打开的 Define Multiple Response Sets 的对话框几乎完全一样。故选择第二种打开方式。但需要指出的是，在两个过程中所做的多重应答集是不兼容的。在 Tables 中做的多重应答集，只能在 Custom Tables 中进行分析，反之亦然。此外，多重应答集的建立是临时性的，关闭 SPSS 后，所建的多重应答集也随之失效。

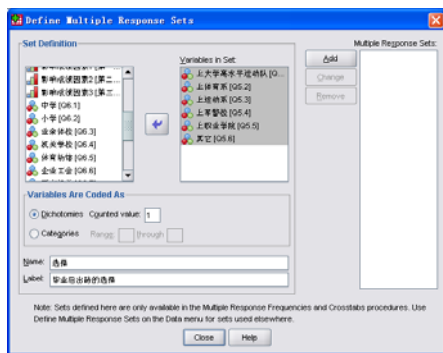


图 2-124 定义多重应答集

③ 在左侧的变量名源框中，选中上大学高水平运动队、上体育系、上运动系、上军

警校、上职业学院、其他变量，用右移箭头将它们移入 Variables in Set 框中。

在 Variables are Coded As 下，选择 Dichotomies 选项，在 Counted value 后框中输入 1，变量取值为 1 表示该变量被选中。

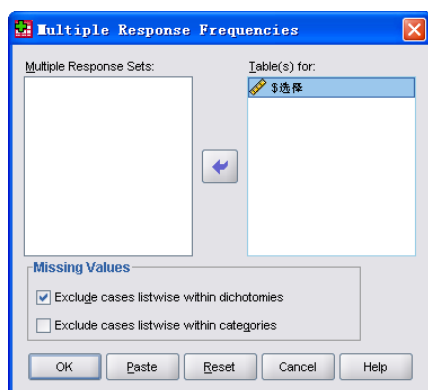


图 2-125 计算多重应答集频数

在 Name 框中输入选择，定义多选题变量集名为选择。在 Label 框中输入毕业后出路的选择，为变量集给出标签。

单击 Add 按钮，正式添加所定义的变量集。

按 Close，关闭 Define Multiple Response Sets 的对话框。定义多重应答集完毕。

下面对所建的多重应答集进行分析。

④ 按 Analyze → Multiple Response → Frequencies，打开 Frequencies 对话框，见图 2-125。

按右移箭头将多重应答集框中的选择变量，移到 Table(s) for 框中。

保持系统默认选择项，单击 OK 按钮运行，在输出窗口中得到频数分布结果，见表 2-58、表 2-59。

表 2-58 样品汇总

Case Summary						
	Oases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
\$选...	186	100.0%	0	.0%	186	100.0%

a. Dichotomy group tabulated at value 1.

表 2-59 多重应答集的频数分布

\$选择 Frequencies				
		Responses		Percent of Cases
		N	Percent	
毕业后出路的选择	上大学高水平运动队	80	23.3%	43.0%
	上体育系	42	12.2%	22.6%
	上运动系	113	32.9%	60.8%
	上军警校	74	21.6%	39.8%
	上职业学院	28	8.2%	15.1%
	其他	6	1.7%	3.2%
Total		343	100.0%	184.4%

a. Dichotomy group tabulated at value 1.

⑤ 结果与讨论

表 2-58 显示了样本含量和有效样本含量均为 186，说明共有 186 名被试者参与了问卷调查的答题工作。

表 2-59 显示了多选题的选择结果。共有 343 人次参与了 6 项选择，186 名被试者中有 80 名选择了毕业后准备上大学高水平运动队，它占 343 人次中的百分比为 23.3%，占 186 人中的比例为 43.0%，占 186 人中的 60.8% 的被试者选择了毕业后上运动系，其余类推。

以下用来分析不同性别之间在这 6 个方面的选择上的分布情况。

⑥ 按 Analyze→Multiple Response→Crosstabs，打开 Crosstabs 对话框，见图 2-126。

在 Multiple Response Sets 框，选中选择变量，按最上面的右移箭头将其移入 Row(s) 框中，在左侧变量名源框中，选择性别变量，并将其用中间的右移箭头移入到 Column(s) 框中。单击 Define Ranges 按钮，打开定义组别的选项卡，见图 2-127。在 Minimum 框中输入 0，在 Maximum 框中输入 1，表明组别用 0、1 区分。按 Continue 返回 Crosstabs 对话框。

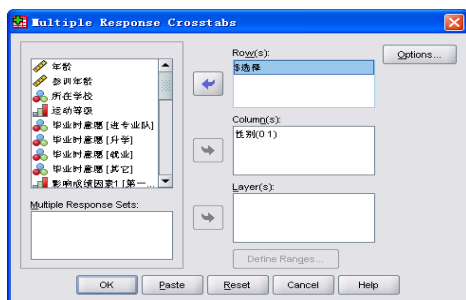


图 2-126 多重应答集交叉表分析

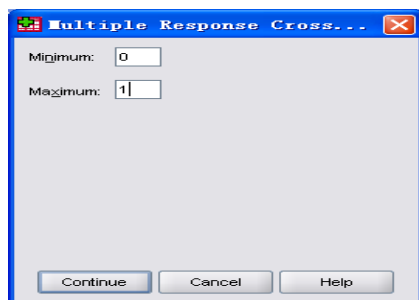


图 2-127 定义组别

单击 Option 按钮，打开 Option 对话框，见图 2-129。在 Cell Percentages 选项中，选择 Row、Column、Total 选择项，要求在输出表中有行、列百分比小计和总计。在 Percentages Based on 中选择 Cases。按样本含量计算百分比。按 Continue 返回 Crosstabs 对话框。

单击 OK 按钮运行，在输出窗口中得到交叉表，见表 2-60、表 2-61。

⑦ 结果与讨论

表 2-60 显示有效样本含量和样本总含量一样都是 186。

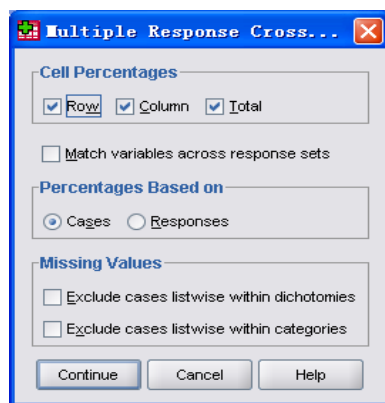


图 2-128 定义对话框

表 2-60 样品汇总

	Case Summary					
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
\$选择*性别	186	100.0%	0	.0%	186	100.0%

表 2-61 多重应答变量选择与性别交叉表

\$选择*性别 Crosstabulation					
					Total
			女	男	
毕业后出路的选择	上大学高水平运动队	Count	25	56	80
		% within \$选择	31.2%	68.8%	
		% within 性别	36.2%	47.0%	
		% of Total	29.6%	13.4%	43.0%
	上体育系	Count	18	24	42
		% within \$选择	42.9%	57.1%	
		% within 性别	26.1%	20.5%	
		% of Total	9.7%	12.9%	22.6%
	上运动系	Count	36	77	113
		% within \$选择	31.9%	68.1%	
		% within 性别	52.2%	65.8%	
		% of Total	19.4%	41.4%	60.8%
	上军警校	Count	26	48	74
		% within \$选择	35.1%	64.9%	
		% within 性别	37.7%	41.0%	
		% of Total	14.0%	25.8%	39.8%
	上职业学院	Count	9	19	28
		% within \$选择	32.1%	67.9%	
		% within 性别	13.0%	16.2%	
		% of Total	4.8%	10.2%	15.1%
	其他	Count	4	2	6
		% within \$选择	66.7%	33.3%	
		% within 性别	5.8%	1.7%	
		% of Total	2.2%	1.1%	3.2%
Total	Count	69	117	186	
	% of Total	37.1%	62.9%	100.0%	

Percentages are based on respondents.

a. Dichotomy grid tabulated at value 1.

从表 2-61 中可见, 选择上大学高水平运动队的人数为 80, 其中女生为 24 人, 男生为 56 人, 各占 80 人次中的比例为 30% 和 70%。从表中倒数第二行中可知, 被试者中, 女生总人数为 69 人, 男生为 117 人, 故在多重应答集选择中的百分比, 与表底最后一行的百分比相比, 不难看出, 男生进行多重选择要比女生更积极一些。因为大于 62.9% 中, 男生中有四项。综合来看, 男生、女生在这 6 项选择上, 有一个共同的特点, 在上运动系选择上都超过了 5 成, 而男生在上大学高水平运动队的选项上, 也将近达到 5 成, 女生则在上军警校选项上将近达到 4 成。

2.7.2 排序题的处理

例 2.52 在对北京市 5 所体校运动员的问卷调查研究中, 同样另有一道题为请在以下影响运动成绩的因素选项中, 按你认为的重要性的大小作出排序选择: _____。

(1) 运动强度 (2) 运动量 (3) 运动持续时间

数据文件存放在 data02-01.sav 中的运动强度、运动量、运动持续时间中。

在 SPSS 中, 同多项题一样, 排序题也可以建成多重应答集, 但利用 SPSS 中现有的对多重应答集的处理功能, 处理结果很难分析。

1. 多重应答集频数分析与变量单独进行频数分析的比较

现用本例为例, 在 SPSS 中建立多重应答集, 对多重应答集进行频数分析, 再用原答案变量单独进行频数分析, 比较两者的关联和区别。

操作步骤如下:

① 在 SPSS 数据编辑窗口中, 打开 data02-01.sav。

② 按 Analyze→Tables→Multiple Response Sets 顺序, 打开 Define Multiple Response Sets 对话框, 见图 2-129。

在左侧变量名源框中, 选中运动强度、运动量、运动持续时间变量, 单击右移箭头将它们移入 Variables in Set 框中。

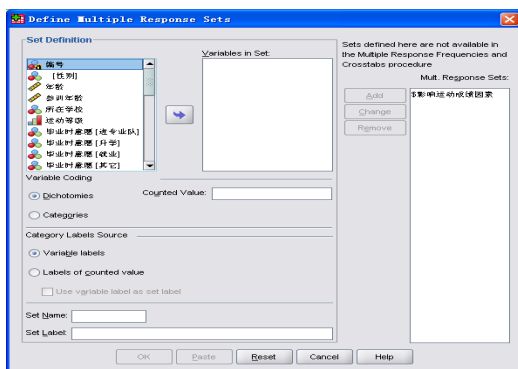


图 2-129 Define Multiple Response Sets 对话框

因为本例有三种可选结果，所以在 Variable Coding 选项中选择 Categories 选项。

在 Set Name 框中输入影响运动成绩因素，单击 OK 按钮，完成多重应答集的建立。

③ 按 Analyze→Tables→Custom Tables 顺序，打开 Custom Tables 对话框，见图 2-92。

在变量名源框中，选择影响运动成绩因素，并将其拖曳到探究窗口的 Columns 上，在 Define 选项中，单击 N% Summary Statistics 按钮，出现 N% Summary Statistics 选项卡，见图 2-130。

在 Statistics 下框中，选择 Column N%，将其拖曳到 Display 下框中，要求输出列百分比。单击左下角 Apply to Selection 按钮，要求在表中应用所作的选择。回到 Custom Tables 对话框。

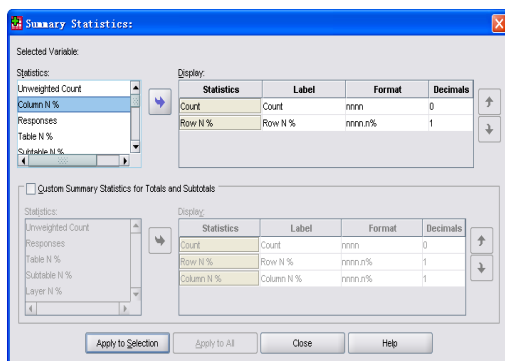


图 2-130 N% Summary Statistics 选项卡

单击 OK 按钮，在输出窗口中得到输出结果，见表 2-62。

表 2-62 影响因素的累计频数分布

\$影响运动成绩因素					
第一重要		第二重要		第三重要	
Count	Column N%	Count	Column N%	Count	Column N%
186	100.0%	186	100.0%	186	100.0%

④ 对原答案变量单独进行频数分析，仍按 Analyze→Tables→Custom Tables 顺序，打开 Custom Tables 对话框，见图 2-92。

从表格探究窗口中移走多重应答集影响运动成绩因素变量。

从左侧变量名源框中，选择运动强度、运动量、运动持续时间，并将其拖曳到 Rows 上，在 Define 选项中，单击 N% Summary Statistics 按钮，出现 N% Summary Statistics 选项卡，见图 2-130。

在 Statistics 下框中，选择 Column N%，将其拖曳到 Display 下框中，要求输出列百分比。单击左下角 Apply to Selection 按钮，要求在表中应用所作的选择。回到 Custom Tables

对话框。

单击 OK 按钮，在输出窗口中得到输出结果，见表 2-62。

⑤ 结果与讨论

从表 2-63 可见，在影响成绩因素中，选择第一位最重要的因素中，119 位选择了运动强度，占总调查人数的百分比为 64.0%，有 36.0% 选择了运动量，而在次重要因素中，有 118 位选择了运动量，占总调查人数的百分比为 63.4%，选择运动持续时间和运动强度的各占 23.1% 和 13.4%，在第三重要因素中，142 位选择了运动持续时间，占总调查人数的百分比为 76.9%，有 22.6% 的运动员选择了运动强度，只有 0.5% 的运动员选择了运动量。由此可见，在接受调查的运动员中，他们对影响运动成绩因素的排位中，大部分倾向于这样的顺序，即运动强度、运动量和运动持续时间。

表 2-63 变量排序的频数分布表

		Count	Column N%
运动强度	第一重要	119	64.0%
	第二重要	25	13.4%
	第三重要	42	22.6%
运动量	第一重要	67	36.0%
	第二重要	118	63.4%
	第三重要	1	.5%
运动持续时间	第一重要	0	.0%
	第二重要	42	23.1%
	第三重要	142	76.9%

表 2-62 显示的是被试者对某一因素在三个位置上的总的被选中的次数，当所有被试者都对这三个因素在三个位置做出选择时，显然每个因素都被选中了 186 次。因此，用总的选中次数并不能够看出因素的重要性的位置。

2. 目前处理排序题的一个常用方法

仍以例 2.52 为例，SPSS 中的操作步骤如下：

(1) 在 SPSS 变量编辑窗口中，建立四个变量，第一个为影响成绩因素（数值型名义测度），第二个为第一重要（数值型尺度变量），第三个为第二重要（数值型尺度变量），第四个为第三重要（数值型尺度变量）。将表 2-63 中的结果按图 2-131 的方式，在 SPSS 数据编辑窗口中建成数据文件，并存放在 data02-39.sav 中。

	影响成绩因素	第一重要	第二重要	第三重要
1	1.00	119.00	25.00	42.00
2	2.00	67.00	118.00	1.00
3	3.00	0.00	43.00	143.00

图 2-131 建立数据文件

(2) 对这三个重要性的位置赋权

一般而言,要准确测定这三个位置的权重是一件很困难的事情。如果能确切测得各位置的权重,也就不再是统计研究的范畴。因为问题提出的本身就是在无法用定量方法解决的前提下需要借助被试者的经验来确定的。因此,对权重的确定依然要靠经验来解决。所以,有许多研究人员对本类问题的研究借助于模糊数学方法来实现。但这不是本书研究的范畴。

显而易见,当所有研究者对这三个因素的重要性位置都有一致的测度时,则判定重要性的测度方法与位置秩之间就产生一致性,自然而然,研究人员就想到用位置秩本身赋权也就不难理解了。

所以,一般直接给这三个位置赋予的权重分别为 1、2、3。

(3) 重要性位置的判定方法

考虑到有的被试者可能只给出所有因素中部分因素的位置秩,所以直接用秩和大小来判定不太合理,因此,可以根据平均秩大小排定重要性位置。

平均秩 = (第一重要 + 第二重要 × 2 + 第三重要 × 3) / (第一重要 + 第二重要 + 第三重要)

由此,在 SPSS 中,可以做如下处理:

(4) 按 Transform → Compute Variable 顺序,打开 Compute Variable 对话框(见图 2-53)。在 Target Variable 的下框中输入目标变量名 平均秩。

在 Numeric Expression 的下框中输入: (第一重要 + 第二重要 * 2 + 第三重要 * 3) / (第一重要 + 第二重要 + 第三重要)。

按 OK 按钮运行,则在原数据文件中生成平均秩新变量,见图 2-132。

	影响成绩因素	第一重要	第二重要	第三重要	平均秩
1	1.00	119.00	25.00	42.00	1.59
2	2.00	67.00	118.00	1.00	1.65
3	3.00	0.00	43.00	143.00	2.77

图 2-132 平均秩计算

(5) 结论

因为运动强度、运动量、运动持续时间的平均秩依次为 1.59、1.65 和 2.77,所以影响成绩因素的重要性排序也依次为:运动强度、运动量、运动持续时间。

3. 排序题的另一种处理方法

对例 2.52 还有另一种排序和检验方法,它可在非参数假设检验中使用多个相关样本中的 Kendall's W 协和系数法来处理。参见第 7 章 7.8.2 中的例 7.23。

第3章 常见的几种概率分布

在自然界中，我们会观察到各种现象，习惯上将其分为两大类，一类是在一定条件组合下，必然发生的现象称为必然发生的确定性现象。如在一个标准大气压下，水煮到 100°C 时会沸腾；在空中无障碍物的情况下，用手上抛的篮球到达最高点后下落。另一类是在一定条件组合下，有多种可能发生的结果，但在每次观察中，发生的结果都不能事先确定的现象称为随机现象。如在重复做抛掷硬币的试验中，我们知道在正常情况下可能的结果只有两个：要么正面朝上，要么反面朝上，但在一次试验中，我们无法预知这次究竟是正面朝上还是反面朝上。虽然，随机现象表现为不确定性，但在大量重复的试验中，还是能观察到某种特定的规律，称作随机现象的统计规律。如蒲丰在做大量的抛掷硬币的试验中发现，虽然在一次试验中，事先无法确定究竟哪一面朝上，但在多次试验后观察到，正面朝上的次数和反面朝上的次数越来越接近。

本章将介绍统计中最基本的概率及其常用的几种分布。

3.1 事件和概率

3.1.1 事件

1. 随机试验

这里所说的试验包括各种科学实验，也包括对某一事物的某一特征的观察或测量。如果一个试验具有以下三个特征，则称该试验为随机试验，即

- (1) 试验可以在相同条件下重复进行；
- (2) 在给定条件下试验结果不只一个；
- (3) 每次试验事先不能准确地预料它出现哪一个结果，但知道可能出现哪些结果。

例如，对挑边器做均匀试验，上抛落地后观察其正面和反面出现的情况，由于该试验可以在相同条件下重复进行，在给定条件下试验结果不只一个，每次试验事先不能准确地预料它出现哪一个结果，但知道可能出现哪些结果，因此，该试验就是随机试验。广义地讲，对任何一个随机现象的观测，均可看成是一个随机试验。

2. 随机事件

随机试验中，在一次试验中每一个可能出现也可能不出现的结果称为随机事件。

例如, 在八名一百米跑运动员参加的决赛中, 运动员的名次将从 1 至 8 名或无名次 9 种可能的结果中取得其中的一种结果, 因此, 1 至 8 名或无名次 9 种可能的结果为随机事件。

(1) 基本事件: 随机试验中不可能再分的结果称为基本事件。上例中 1 至 8 名的每一个结果及无名次均为基本事件, 因为它们不能再分。

(2) 复合事件: 由若干个基本事件组合而成的事件称复合事件。例如, 上例中, “某运动员取得前三名名次”这一事件就是复合事件, 因为“取得前三名”这一事件是由“取得第一名”、“取得第二名”或“取得第三名”这三个基本事件组合而成的事件。

3. 必然事件

随机试验中必然会出现的结果称必然事件。如在正常状态下, 完成比赛的运动员必定有他相应的成绩就是必然事件。

4. 不可能事件

随机试验中不可能出现的结果称不可能事件。如在正常状态下, 未完成比赛的运动员有成绩就是不可能事件。

无论是必然事件, 还是不可能事件, 由于它们在事前我们都能预先知道它的结果, 因此, 它是经典数学研究的范畴。

3.1.2 事件之间的关系和运算

1. 包含

若事件 A 的发生必然导致事件 B 的发生, 则称事件 B 包含事件 A , 记作 $A \subset B$ 或 $B \supset A$ 。

在体育比赛中, 若令 A 表示{取得 3 到 4 名}的事件, B 表示{取得前六名}的事件, 则 A 的发生必然导致 B 的发生, 因而 $A \subset B$ 。

2. 相等

若 $A \supset B$ 且 $B \supset A$ 则称 $A = B$ 。

如令 A 表示{月收入不低于 1000 元}的事件, B 表示{月收入至少 1000 元}的事件, 则显然有 $A = B$ 。

3. 和

若事件 A 和事件 B 至少一个发生时就使得事件 C 发生, 则称事件 C 为 A 与 B 的和, 简称为和, 记作 $C = A \cup B$ 。

如令 A 表示{取得前 4 名}的事件, B 表示{取得 4 到 6 名}的事件, C 表示{取得前 6 名}的事件, 比赛结果为第 3 名 (A 发生), 或第 5 名 (B 发生), 则事件 C 也发生。

4. 积

若事件 A 和事件 B 同时发生事件 C 才发生, 称事件 C 为事件 A 与事件 B 的积事件,

简称为积, 记作 $C=A \cap B$ 。

如令 A 表示{取得前4名}的事件, B 表示{取得4到6名}的事件, 由于 $C=A \cap B$ 因此, C 为{取得第4名}的事件。

5. 差

称事件 A 发生但事件 B 不发生的事件为事件 A 减事件 B 的差事件, 简称为差, 记作 $A-B$ 。

上述和例子中 $A-B$ 表示{取得前3名}的事件。

6. 互斥

若事件 A 和事件 B 不能同时发生, 则称 A 与 B 互斥或互不相容。互斥包括非此即彼的情形, 但互斥并非全是非此即彼, 事件的关系满足 $A \cap B = \emptyset$ (不可能事件)。

如在抛掷均匀硬币的试验中, 令 A 表示{正面朝上}的事件, B 表示{反面朝上}的事件, 则 A 和 B 互斥, 试验的结果表现为非此即彼。

而在掷骰子的试验中, 令 A 表示{出现小于等于3点}的事件, B 表示{出现大于等于5点}的事件, 则 A 和 B 用样互斥, 但试验的结果并非为非此即彼, 还可能{出现4点}的事件。

7. 对立

称事件 A 不发生就发生的事件为 A 的对立事件, 记为 \bar{A} 。在一次试验中, 事件 A 与事件 \bar{A} 的发生非此即彼, 故 $A \cup \bar{A} = \Omega$ (必然事件), 并且 $A \cap \bar{A} = \emptyset$ 。

非此即彼的互斥事件即为对立事件。

8. 独立

事件 A 的发生不影响事件 B 发生的概率, 则称事件 A 与事件 B 相互独立, 反之亦然, A 与 B 是一对彼此独立的事件。

根据随机试验的概念可知, 随机试验的本质是重复独立试验, 观察一名篮球运动员的投篮结果就是一次试验, 观察 n 次投篮结果就是 n 次重复独立试验。

9. 完备事件系

若 n 个事件 A_1, A_2, \dots, A_n 两两互斥, 且满足

$$P(A_1 + A_2 + \dots + A_n) = P(A_1) + P(A_2) + \dots + P(A_n) = 1$$

则称该 n 个事件为一个完备事件系。这两个条件缺一不可。

3.1.3 事件的频率和概率

随机事件发生与否在一次试验中, 我们事先谁都无法肯定, 但随机事件发生在多次试验中有其统计规律性。

3.1.3.1 事件的频率

对于随机事件 A，若在 n 次独立重复的试验中，出现 m 次，就称 m 为事件 A 在 n 次独立重复的试验中出现的频数，称 m/n 为事件 A 的频率。记为： $f(A)=m/n$ 。

所谓重复，是指每次的试验条件要相同，而独立指的是前一次的试验结果不影响后一次的试验结果发生的概率。

例如，某名篮球运动员在罚球区投篮 20 次，投中 18 次，在理论上我们假定该篮球运动员投篮技术已达到动力定型，且训练有素，则对其而言在罚球区投篮 20 次，就是做了 20 次的独立重复的试验。所以，该名篮球运动员的投篮命中的频率为

$$f=m/n=18/20=0.9$$

因为 m 的取值范围为 $[0, n]$ 之间的整数，所以，对任意一个随机事件而言，总有

$$0 \leq f(A) \leq 1$$

3.1.3.2 事件的概率

在随机试验中，每个基本事件称为样本点，样本点的全体称样本空间，即必然事件，记作 Ω 。由此，随机事件就是样本空间中的子集合。

用以度量事件发生可能性大小的数值称作事件的概率。通常用大写字母如 A、B 等表示相应的事件，而其概率通常用 $P(A)$ 、 $P(B)$ 等表示。

事件的概率有以下性质：

- (1) 对于任意一个事件 A，有 $0 \leq P(A) \leq 1$ ；
- (2) 对于必然事件 Ω ，有 $P(\Omega)=1$ ；
- (3) 对于不可能事件 ϕ ，有 $P(\phi)=0$ 。

1. 概率的古典定义

(1) 古典概型

有以下特征的随机试验模型，称为古典概型（概率模型的简称）。

- ① 试验的样本空间只有有限个样本点；
- ② 试验中每个样本点出现的可能性相同。

(2) 概率的古典定义

设基本的样本空间为 $\{e_1, e_2, \dots, e_n\}$ ，事件 $A = e_{k_1}, e_{k_2}, \dots, e_{k_r}$ ，其中 k_r 为 $1, 2, \dots, n$ 中任意 r 个不同的数，则事件 A 的定义为

$$P(A) = \frac{r}{n}$$

称这种概率定义为概率的古典定义。

例 3.1 将一枚均匀的硬币连续抛掷 3 次，求恰有一次正面向上的概率。

解：该试验是一个古典概型。样本空间的基本事件的总数为 $n = C_2^1 \cdot C_2^1 \cdot C_2^1 = 8$ ，令 A

表示{恰有一次出现正面向上}的事件, 则 A 包含的基本事件数为 $C_3^1 = 3$, 故恰有一次正面向上的概率 $P(A) = 3/8$ 。

例 3.2 设有一批产品共 100 件, 其中有 5 件次品, 现从中任取 50 件, 问恰有 2 件是次品的概率是多少?

解: 从 100 件产品中任取 50 件, 共有 C_{100}^{50} 个结果, 每个结果就是一个基本事件, 显然这些事件是一个等概率基本事件组, 即基本的样本空间点, 令 A 表示{恰有 2 件是次品}的事件, 在取出的 50 件产品中, 要求恰有 2 件是次品, 即有 48 件是从 95 件正品中抽取的, 有 2 件是从 5 件次品中抽取的, 它共有 $C_{95}^{48}C_5^2$ 个基本事件。

故 $P(A) = C_{95}^{48}C_5^2 / C_{100}^{50} = 0.32$ 。

2. 随机事件概率的统计学定义

在一定的条件组合下, 进行大量的重复独立试验, 随着试验次数的增多, 某一随机事件 A 发生的频率总是围绕着某个固定值 P 而波动, 这个固定值 P 就称为随机事件 A 发生的概率。记作 $P(A)$ 。

例 3.3 在抛掷均匀硬币的实验中, 得到如下统计结果, 见表 3-1。求正面出现朝上的概率。

表 3-1 抛掷均匀硬币的实验结果

试验者	抛硬币次数	正面出现的次数	频率
德·摩尔根	2048	1061	0.518
蒲丰	4040	2048	0.5069
皮尔逊	12000	6019	0.5016
皮尔逊	24000	12012	0.5005
维尼	30000	14994	0.4998

解: 在表 3.1 中, 可以发现, 随着试验次数的增多, 频率开始围绕着 0.5 而波动, 所以, 根据随机事件概率的统计学定义可知, 0.5 就称为抛掷均匀硬币“出现正面朝上”这个随机事件发生的概率。

一般而言, 概率越接近 0, 表示事件发生的可能性越小; 概率越接近 1, 表示事件发生的可能性越大。

3. 条件概率

在 B 事件已发生的前提下, A 事件发生的概率, 称做事件 A 的条件概率, 记为 $P(A|B)$ 。

例 3.4 设盒中装有 16 个球, 其中 6 个是玻璃球, 另外 10 个是木质球。玻璃球中有 2 个是红色的, 4 个是蓝色的; 木质球中有 3 个是红色的, 7 个是蓝色的。现从中任取一

个, 记 $A=\{\text{取到蓝球}\}$, $B=\{\text{取到玻璃球}\}$, 设该球为蓝球, 则该球为玻璃球的概率为多大?

解: 由题意可知, 蓝球共有 $4+7=11$ 个, 而 11 个蓝球中共有 4 个是玻璃球。根据古典概型可得: $P(B|A)=4/11$ 。

此外, 本例中, 由于已知 $P(A)=\frac{11}{16}$, $P(AB)=\frac{4}{16}$, 故根据乘法公式: $P(AB)=P(A)P(B|A)$, 可得 $P(B|A)=P(AB)/P(A)=4/11$ 。

3.2 随机变量和概率分布

3.2.1 随机变量

1. 随机变量

概率论和数理统计都是从数量的侧面来研究随机现象的统计规律性, 建立起一系列公式和定理, 由此对实际问题展开研究和分析, 其中随机变量的概率分布就是用来反映随机变量的内在规律性的。

设随机试验 E 的样本空间为 Ω , 如果对于每一个样本点 e , 都有一个实数 X 与之对应, 则称 X 为随机变量。

随机变量 X 是一个以基本事件 e 为自变量的取实值的函数。 X 的所有可能值是知道的, 但在一次确定的试验结束之前, 事先无法确定 X 究竟取什么值, 所以 X 的取值具有随机性。由于试验的各个结果的出现有一定的概率, 所以随机变量的取值也有一定的概率规律。

2. 随机变量的分布函数

(1) 定义: 设 X 是一个随机变量, x 是任意实数, 称 $P\{X \leq x\}$ 为 X 的分布函数记为 $F(x)$

$$F\{x\} = P\{X \leq x\}$$

这里, 分布函数 $F(x)$ 是 x 的一个普通的实函数。

(2) 分布函数的性质:

$$0 \leq F(x) \leq 1$$

$F(x)$ 是 x 的单调非降函数。

$$F(-\infty) = \lim_{n \rightarrow -\infty} F(x) = 0, \quad F(+\infty) = \lim_{n \rightarrow \infty} F(x) = 1$$

$F(x)$ 是右连续函数, 即满足 $\lim_{x \rightarrow x_0^+} F(x) = F(x_0)$ 。

(3) 应用：许多重要的随机事件概率的计算。

$$P\{a \leq X \leq b\} = F(b) - F(a)$$

$$P\{X = a\} = F(a) - F(a-0) = F(a) - \lim_{x \rightarrow a} F(x)$$

$$P\{X > a\} = 1 - F(a)$$

$$P(X \geq a) = 1 - F(a-0)$$

3. 随机变量的分类

随机变量分为离散型（可以把 X 可能取的值一一列举出来，如计数资料）和连续型（不可以把 X 可能取的值一一列举出来，如计量资料）两种，随机变量的概率分布也分为离散型和连续型两种。

3.2.2 离散型随机变量的概率分布

3.2.2.1 概率分布

1. 定义：设离散型随机变量可能取的值为： $x_1, x_2, \dots, x_k, \dots$ ，随机变量 X 取这些值的概率为

$$p_k = P\{X = x_k\} \quad (k = 1, 2, \dots)$$

则上式称为 X 的概率分布。

2. 概率分布表

把 X 可能取的值及相应的概率列成的表，称为概率分布表，见表 3-2。

表 3-2 概率分布表

X	x_1	x_2	\dots	x_k	\dots
p	p_1	p_2	\dots	p_k	\dots

3. 期望

称和数 $\sum_k x_k p_k = x_1 p_1 + x_2 p_2 + \dots + x_k p_k + \dots$ 为随机变量 X 的期望（又称数学期望），

记作 $E(X)$ 。

4. 方差

称和数 $\sum_k [x_k - E(X)]^2 p_k$ 为随机变量 X 的方差。记作 $D(X)$ 。

3.2.2.2 二点分布

1. 定义

若随机变量 X 的分布为

$$p\{X=1\}=p \quad (0 < p < 1)$$

$$p\{X=0\}=q=1-p$$

则称 X 服从二点分布。它又称伯努利分布。

2. 期望与方差

二点分布的期望为： p 。

二点分布的方差为： pq 。

3. 应用背景

当一组条件下只有两个可能结果，且都有正概率，即事件 A 发生的概率为 $p(0 < p < 1)$ ，不发生的概率为 $q=1-p$ ，则在单次试验中，这个随机变量服从二点分布。

4. SPSS 中与二点分布有关的函数

PDF.BERNOULLI(quant, prob)函数，它计算得到给定的参数 prob 下，等于 quant 时的二点分布的概率。

CDF.BERNOULLI(quant, prob)函数，它计算得到给定的参数 prob 下，小于等于 quant 时的二点分布的累积概率。在二点分布中，quant 的取值为 0 和 1。

5. 实例及其计算

例 3.5 在豌豆的红花纯合基因型与白花纯合基因型的杂交试验中，得到杂交品种出现白花种子的概率为 0.25，出现红花种子的概率为 0.75，现从其杂交品种中随机抽取一粒杂交种子，问该种子为红花种子的概率是多少？

由于该杂交试验中只有白、红两种结果，又符合任意抽取一粒种子是“白”或“红”的模型，所以，这是一个二点分布的情形，根据二点分布可知，该种子为红花种子的概率是 0.75。

在 SPSS 中，可用下述步骤来处理此类问题：

① 在 SPSS 数据编辑窗口，建立数据文件，见 data03-01.sav。

② 按 Transform→Compute Variable 顺序，打开 Compute Variable 对话框（见图 2-53）。在 Target Variable 框中，输入目标变量名为该粒为红花种子的概率。

③ 在 Numeric Expression 框中，输入“PDF.BERNOULLI(0,白花种子的概率)”。

④ 单击 OK 按钮运行，则在数据编辑窗口工作的数据文件中，出现该粒为红花种子的概率的新变量及其值。见图 3-1。

即该种子为红花种子的概率是 0.75。

	白花种子的概率	该粒为红花种子的概率
1	0.25	0.75

图 3-1 该粒为红花种子的概率

3.2.2.3 二项分布

1. 定义

如果随机变量 X 的分布如下

$$P\{X = k\} = C_n^k p^k q^{n-k} \quad (k = 0, 1, 2, \dots, n) \\ (0 < p < 1, q = 1 - p)$$

则称 X 服从二项分布，或记为： $X \sim B(n, p)$ 。

当 $np = \lambda$ 为常量时，二项分布即为泊松分布，所以当 n 很大，而 p 较小时，二项分布可以用 3.2.1.4 中介绍的泊松分布来近似。

此外，当 $n \rightarrow \infty$ 时，理论上已经证明，二项分布的极限分布为正态分布。

2. 期望与方差

二项分布的期望为： np

二项分布的方差为： npq

3. 应用背景

单次试验中，事件 A 发生的概率为 p ($0 < p < 1$)，不发生的概率为 $q = 1 - p$ ，则在 n 次独立试验中， A 发生的次数 X 这个随机变量服从二项分布。

4. SPSS 中与二项分布相关的函数

PDF.BINOM(*quant*,*n*,*prob*)函数，其中的参数分别为： n 是总的试验次数，*quant* 是某事件出现的次数，*prob* 是该事件出现的概率。用它可计算得到的是 n 次试验中，事件出现的概率值为 *prob* 时，某事件出现的次数等于 *quant* 时的概率。

CDF.BINOM(*quant*,*n*,*prob*) 函数是二项分布的累积函数，参数的含义同上。它计算得到的是 n 次试验中，事件出现的概率值为 *prob* 时，某事件出现的次数小于等于 *quant* 的累积概率。

5. 实例及其计算

例 3.6 某机器生产的产品的次品率为 0.005，在它生产 1000 件的产品中，

- (1) 有 4 件次品的概率是多少？
- (2) 次品至少为 2 件的概率是多少？
- (3) 在 1000 件产品中，有几件次品可能性最大？

在 SPSS 中，处理本类问题的步骤如下：

① 在 SPSS 数据编辑窗口，建立数据文件，见 data03-02.sav。

② 按 Transform→Compute Variable 顺序，打开 Compute Variable 对话框(见图 2-53)。在 Target Variable 框中，输入目标变量名为 *出现四件次品的概率*。

③ 在 Numeric Expression 框中，输入“PDF.BINOM(4,试验总次数,次品出现的概率)”或“CDF.BINOM(4,试验总次数,次品出现的概率)–CDF.BINOM(3,试验总次数,次品出现的

概率)”。

- ④ 单击 OK 按钮运行，则在数据编辑窗口工作的数据文件中，出现一列变量名为出现四件次品的概率的值。见图 3-2。

	试验总次数	次品出现的概率	出现四件次品的概率
1	1000.00	0.005	0.18
2			

图 3-2 例 3.6 中 (1) 的答案

即有 4 件次品的概率是 0.18。

- ⑤ 按 Transform→Compute Variable 顺序，打开 Compute Variable 对话框。在 Target Variable 框中，输入目标变量名为出现次品至少为 2 件的概率。在 Numeric Expression 框中，输入 “1-CDF.BINOM(1,试验总次数,次品出现的概率)”。

- ⑥ 单击 OK 按钮运行，则在数据编辑窗口工作的数据文件中，出现一列变量名为出现次品至少 2 件的概率。见图 3-3。

	试验总次数	次品出现的概率	出现四件次品的概率	出现次品至少为2件的概率
1	1000.00	0.005	0.18	0.96
2				

图 3-3 例 3.6 中 (2) 的答案

即至少有件次品的概率是 0.96。

- ⑦ 求有几件次品可能性最大

为回答“有几件次品可能性最大”的问题，我们首先设有次品 k_0 时可能性最大，则可得联立不等式

$$\begin{cases} \frac{P_n(k_0)}{P_n(k_0-1)} \geq 1 \\ \frac{P_n(k_0)}{P_n(k_0+1)} \leq 1 \end{cases}$$

解上述联立不等式得

$$k_0 = \begin{cases} (n+1)P-1 \text{ 或 } (n+1)P, & \text{当 } (n+1)P \text{ 是整数时} \\ [(n+1)P] & , \text{当 } (n+1)P \text{ 不是整数时} \end{cases}$$

式中 $[(n+1)P]$ 表示对 $(n+1)P$ 的值取整数。

具体做法同⑤、⑥。在 Target Variable 框中，输入目标变量名为出现可能性最大的次品数。在 Numeric Expression 框中，输入 “TRUNC((试验总次数+1)*次品出现的概率)”，TRUNC(numexpr)是截尾函数。单击 OK 按钮运行，则在数据编辑窗口中，出现一列变量名为出现可能性最大的次品数值。见图 3-4。

	试验总次数	次品出现的概率	出现四件次品的概率	出现次品至少为2件的概率	出现可能性最大的次品数
1	1000.00	0.005	0.18	0.96	5.00
2					

图 3-4 例 3.6 中 (3) 的答案

即 1000 件产品中最有可能出现 5 件次品。

例 3.7 设某射击运动员命中 10 环的概率为 0.99, 现在射击比赛中, 共射击 100 发, 求至少有 95 发打中 10 环的概率。

【题析】 至少有 95 发打中 10 环的概率等价于用 1 减去 94 发以下打中 10 环的概率。

因此, 仿例 3.5 的做法, 先在 SPSS 中建立数据文件, 见 data03-03.sav。在 Compute Variable 对话框的 Target Variable 框中, 输入目标变量名为至少有 95 发打中 10 环的概率。再在 Numeric Expression 框中, 输入“1-CDF.BINOM(94, 射击 100 次, 射中 10 环概率)”。单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现一列变量名为至少有 95 发打中 10 环的概率 (将其小数位数设定为 6 位)。见图 3-5 的第三列。

故至少有 95 发打中 10 环的概率为 0.999465。

	射击100次	射中10环概率	至少有95发打中10环的概率
1	100.00	0.99	0.999465
2			

图 3-5 例 3.7 数据文件与计算结果

例 3.8 某校上报的体育达标率为 0.75, 现从该校随机抽测 200 名学生, 结果有 147 名学生达到体育锻炼标准, 试问该校上报的体育达标率是否是虚假的?

本例换一种角度, 用假设检验的思路来回答它。

$H_0: p = p_0$ (即抽样达标率与该校上报的达标率相同)

在原假设成立的前提下, 仿例 3.6 (2) 的做法, 可得从 200 人中, 抽得 147 人以下达标的概率为 0.3375。

由于在抽样达标率与该校上报的达标率相同的原假设下, 观测的显著性水平为 0.3375 大于 0.05, 所以, 现有的证据不支持该校上报的达标率是虚假的判断。

在 SPSS 中, 还可以用第 7 章非参数假设检验中的二项分布检验法来完成本例的检验。参见第 7 章。

例 3.9 在保险公司里有 2500 个同一年龄和同社会阶层的人参加了保险。在一年里每个人死亡的概率为 0.002, 每个参加保险的人在一月一日付 12 元保险费, 而在死亡时家属可从公司里领 2000 元, 问 (a) 保险公司亏本的概率是多少? (b) 公司获利不少于 10000 元的概率是多少?

【题析】

在一年中, 保险公司的收入为 $2500 \times 12 = 30000$ (元), 若一年中死亡的人数为 x 人,

则保险公司应付出 $2000x$ 元。

保险公司亏损的条件为 $2000x > 30000$ ，即 $x > 15$ （人），也就是当死亡人数超过 15 人时，保险公司亏本。所以，本例问题（a）可以转化为死亡人数超过 15 人的概率。

而（b）公司获利不少于 10000 元，等价于 $30000 - 2000x \geq 10000$ ，即 $x \leq 10$ （人），故“获利不少于 10000 元”等价于“一年中死亡的人数不超过 10 人”。

本例在 SPSS 中的解题步骤如下：

仿例 3.5 的做法，先在 SPSS 中建立数据文件，见 data03-04.sav，数据文件中的变量均为常数变量。在 Compute Variable 对话框的 Target Variable 框中，输入目标变量名为 亏本的概率。再在 Numeric Expression 框中，输入“1-CDF.BINOM（死亡人数 15 人以上，总人数，死亡率）”。单击 OK 按钮运行，则在数据编辑窗口工作的数据文件中，出现一列变量名为 亏本的概率（将其小数位数设定为 6 位）。见图 3-6 的第五列。

	总人数	死亡率	死亡人数15人以上	死亡人数不超过10人	亏本的概率	获利不少于10000元的概率
1	2500.00	0.0020	15.00	10.00	0.000067	0.9864

图 3-6 例 3.9 数据文件与计算结果

再在 Compute Variable 对话框的 Target Variable 框中，输入目标变量名为 获利/不少于 10000 元的概率。而在 Numeric Expression 框中，输入“CDF.BINOM(死亡人数不超过 10 人,总人数,死亡率)”。单击 OK 按钮运行，则在数据编辑窗口工作的数据文件中，出现一列变量名为 亏本的概率（将其小数位数设定为 4 位）。见图 3-6 的第六列。

因此，（a）保险公司亏本的概率是 0.000067，（b）公司获利不少于 10000 元的概率是 0.9864。

3.2.2.4 泊松（Poisson）分布

1. 定义

如果随机变量 X 的分布如下

$$P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda} \quad (k = 0, 1, 2, \dots; \lambda > 0)$$

其中： $\lambda = np$ (常数), $n \rightarrow \infty$ 。

2. 期望与方差

泊松分布的期望为： λ

泊松分布的方差为： λ

3. 应用背景

放射性物质分裂时放出的 α 粒子的个数、电子的散弹效应、电话交换台在规定时间内间隔内的呼叫次数、来到某公共汽车站的乘客数等都近似服从泊松分布。

4. SPSS 中与泊松分布相关的函数

PDF.POISSON(*quant*, *mean*)函数, 其中的参数分别为: *quant* 是某事件出现的次数, *mean* 是该事件出现的数学期望, 即 λ 值。用它可计算在给定 *mean* 时某事件出现的次数所对应的概率值。

CDF.POISSON(*quant*, *mean*)函数是泊松分布的累积函数, 其中的参数分别为: *quant* 是某事件出现的次数, *mean* 是该事件出现的数学期望, 即 λ 值。用它计算得到的是在给定的 λ 下, 某事件出现的次数小于等于 *quant* 的累积概率。

5. 实例及其计算

例 3.10 Rutherford 和 Geiger 观察了放射性物质放出的 α 粒子个数情况, 每次观察 7.5 秒, 一共做了 2608 次试验, 总共观察到 10094 个 α 粒子, 结果见表 3-3。

表 3-3 2608 次试验中放射性物质的分布情况

放射性粒子数 X	0	1	2	3	4	5	6	7	8	9	10 以上
观察到次数	57	203	383	525	532	408	273	139	45	27	16

试问 (1) 放射性物质在一段时间内放射的 α 粒子数 X 是否服从泊松分布?

(2) 如它服从泊松分布, 则计算放射性物质放出的 α 粒子个数期望出现的理论次数?

(3) 求一次观察中出现粒子数不大于 4 的概率?

在 SPSS 中, 通过以下的步骤和途径可以完成本例的检验:

① 在 SPSS 数据编辑窗口中, 建立数据文件。具体内容参见 Data03-05.sav。

② 用观察到次数作为权重变量进行加权处理。

③ 按 Analyze \rightarrow Nonparametric Tests \rightarrow 1-Sample K-S 顺序, 打开 One-Sample Kolmogorov-Smirnov Test 对话框, 见图 3-7。

在左边变量名源框中, 选择放射粒子数变量将其移入到 Test Variable List 框中, 在 Test Distribution 选项中只选择 Poisson 选项。

有关 K-S 检验方面的详细介绍, 参见第 7 章中的 K-S 检验方面的相关内容。

④ 单击 OK 按钮执行, 在输出窗口中, 得到图 3-8 所示的输出结果。

⑤ 结果分析

从图 3-8 可见, 放射性粒子数服从泊松分布的原假设成立的概率为 0.850, 所以不拒绝原设计。故可认为放射性粒子数服从泊松分布。

⑥ 求放射性物质放出的 α 粒子个数期望出现的理论次数

由于放射性粒子数服从泊松分布, 故 $\lambda = 10094 / 2608 = 3.87$ (见图 3-8, Mean=3.87)。

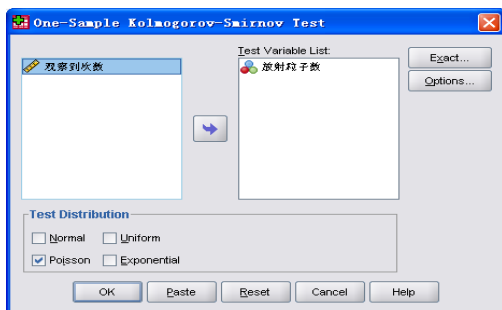


图 3-7 One-Sample K-S Test 对话框

One-Sample Kolmogorov-Smirnov Test			放射粒子数
N			2608
Poisson Parameter ^a	Mean		3.87
Most Extreme Differences	Absolute		.012
	Positive		.010
	Negative		-.012
Kolmogorov-Smirnov Z			.611
Asymp. Sig. (2-tailed)			.850

a. Test distribution is Poisson.

图 3-8 泊松分布检验结果

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 *理论期望次数*。

在 Numeric Expression 框中, 输入 “2608*PDF.POISSON(放射粒子数,10094/2608)”。

单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现一列变量名为 *理论期望次数*。见图 3-9。

⑦ 求一次观察中出现粒子数不大于 4 的概率

首先, 在 SPSS 数据编辑窗口中, 建立数据文件。具体内容参见 Data03-06.sav。

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框, 在 Target Variable 框中, 输入目标变量名为 *出现不大于 4 个粒子的概率*。

在 Numeric Expression 框中, 输入 “CDF.POISSON(4,总的观察到的粒子数/总的试验次数)”。

单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现一列变量名为 *出现不大于 4 个粒子的概率*。见图 3-10。

	放射粒子数	观察到次数	理论期望次数
1	0.00	57.00	54.38
2	1.00	203.00	210.46
3	2.00	383.00	407.28
4	3.00	525.00	525.45
5	4.00	532.00	508.42
6	5.00	408.00	393.56
7	6.00	273.00	253.87
8	7.00	139.00	140.37
9	8.00	45.00	67.91
10	9.00	27.00	29.20
11	10.00	16.00	11.30

图 3-9 理论期望次数计算值

	总的观察到的 粒子数	总的试验次数	出现不大于4个粒子的 概率
1	10094.00	2608.00	0.65

图 3-10 一次观察中出现粒子数不大于 4 的概率

结果: 在一次观察中出现粒子数不大于 4 的概率为 0.65。

3.2.3 连续型随机变量的分布

3.2.3.1 概率密度函数

1. 定义: 对于随机变量 X , 如果存在非负可积函数 $p(x)(-\infty < x < +\infty)$, 使对任意的 $a, b (a < b)$ 都有

$$p\{a < X < b\} = \int_a^b p(x) dx$$

则称 X 为连续型随机变量, 称 $p(x)$ 为 X 的概率密度函数 (简称概率密度)。

2. 性质

作为概率密度函数, 不难推知有以下性质

$$\int_{-\infty}^{\infty} p(x) dx = 1$$

3. 期望

称 $\int_{-\infty}^{+\infty} xp(x) dx$ 为 X 的期望, 记作 $E(X)$ 。

4. 方差

称 $\int_{-\infty}^{+\infty} (x - E(x))^2 p(x) dx$ 为 X 的方差, 记作 $D(X)$ 。可以推知: $D(X) = E[X - E(X)]^2$ 。

3.2.3.2 均匀 (Uniform) 分布

1. 定义

如果连续型随机变量 X 的概率密度为

$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{其他} \end{cases}$$

则称 X 在区间 (a, b) 上服从参数为 a 和 b 的均匀分布。记作 $X \sim U(a, b)$ 。

由此可得, X 的分布函数为

$$F(x) = \int_{-\infty}^x f(x) dx = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x < b \\ 1, & x \leq b \end{cases}$$

2. 期望与方差

均匀分布的期望为: $\frac{a+b}{2}$

均匀分布的方差为：
$$\frac{(b-a)^2}{12}$$

3. 应用背景

在计算机上定点计算中的舍入误差一般认为服从均匀分布。

4. SPSS 中与均匀分布相关的函数

IDF.UNIFORM(*prob*, *min*, *max*)函数, 其中的参数分别为: *prob* 是累积概率值, *min* 是给定区间的最小值, *max* 是给定区间的最大值。用它可计算得到在由 *min* 和 *max* 决定的区间中给定累积概率 *prob* 时所对应的随机变量 *X* 的值。

CDF.UNIFORM(*quant*, *min*, *max*)函数是均匀分布的累积函数, 它用来计算在给定的区间中, 随机变量值小于 *quant* 的累积概率。

PDF.UNIFORM(*quant*, *min*, *max*)函数是用来计算在给定的区间中, 随机变量 *X* 值在 *quant* 处的概率密度值。

5. 实例及其计算

例 3.11 设公共汽车站每隔 10 分钟有一辆公共汽车通过, 则在任一时刻到站乘客的候车时间 *X* (单位: 分钟) 在区间 (0, 10) 上服从均匀分布, 求 $P(1 < x < 3)$ 。如果要有 80% 的把握能坐上车, 乘客至少在车站要等多少分钟?

在 SPSS 中, 通过以下的步骤和途径可以完成本例的计算:

① 在 SPSS 数据编辑窗口中, 建立包括最小值、最大值两个变量的数据文件。具体内容参见 Data03-07.sav。

② 按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框, 在 Target Variable 框中, 输入目标变量名为在 1 到 3 分钟之间的概率。

在 Numeric Expression 框中, 输入 “CDF.UNIFORM(3,最小值,最大值)-CDF.UNIFORM(1,最小值,最大值)”。

单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现一系列变量名为在 1 到 3 分钟之间的概率。见图 3-11。

	最小值	最大值	在1到3分钟之间的概率
1	0.00	10.00	0.20

图 3-11 乘到车的概率

结果, 在 1 到 3 分钟之间的概率为 0.20。

③ 按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框, 在 Target Variable 框中, 输入目标变量名为乘客至少在车站要等车的时间。

在 Numeric Expression 框中, 输入 “IDF.UNIFORM(0.8,最小值, 最大值)”。

单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现一列变量名为 *乘客至少在车站要等车的时间*。见图 3-12。

	最小值	最大值	在1到3分钟之间的概率	乘客至少在车站要等车的时间
1	0.00	10.00	0.20	8.00

图 3-12 在车站候车的时间

④ 结果

乘客至少在车站要等车的时间为 8 分钟。

3.2.3.3 指数 (Exponential) 分布

1. 定义

如果随机变量 X 的概率密度为

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

其中 $\lambda > 0$ 为常数, 则称 X 服从参数为 λ 的指数分布。记作 $X \sim e(\lambda)$ 。

其分布函数为

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

2. 期望与方差

指数分布的期望为: $\frac{1}{\lambda}$

指数分布的方差为: $\frac{1}{\lambda^2}$

3. 应用背景

产品寿命、动物寿命以及某些稀有事件发生的时间间隔等一般认为其服从指数分布。

4. SPSS 中与指数分布相关的函数

CDF.EXP(*quant*, *scale*)函数, 在给定尺度参数 *scale* 下, 计算值小于 *quant* 的指数分布的累积概率。

IDF.EXP(*p*, *scale*)函数, 在给定尺度参数 *scale* 下, 计算累积概率为 *p* 时的随机变量 X 的值。

PDF.EXP(*quant*, *shape*)函数, 在给定形状参数 *shape* 下, 计算随机变量 X 值在 *quant* 处的指数分布概率密度值。

5. 实例及其计算

例 3.12 从某个产品寿命的试验中, 得随机变量寿命的概率密度为

$$f(x) = \begin{cases} 0.05e^{-0.05x}, & x > 0 \\ 0, & x < 0 \end{cases}$$

(1) 求 $P|10 < x \leq 20|$, 即求产品寿命 x 大于 10 且小于等于 20 的概率。

(2) 如果要使概率 $P|X > x| < 0.1$, 则 x 应在什么范围内取值?

在 SPSS 中, 通过以下的步骤和途径可以完成本例的计算:

① 在 SPSS 数据编辑窗口中, 建立只包括尺度参数一个变量的数据文件。具体内容参见 Data03-08.sav。

② 按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框, 在 Target Variable 框中, 输入目标变量名为在 10 到 20 之间的概率。

在 Numeric Expression 框中, 输入 “CDF.EXP(20,尺度参数)-CDF.EXP(10,尺度参数)”。

单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现一系列变量名为在 10 到 20 之间的概率。见图 3-13。

	尺度参数	在10到20之间的概率	X大于x的概率
1	0.05	0.2387	46.05

图 3-13 所求的概率及 X 值

因此, $P|10 < x \leq 20| = 0.2387$ 。

③ 按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框, 在 Target Variable 框中, 输入目标变量名为 X 大于 x 的概率。

在 Numeric Expression 框中, 输入 “IDF.EXP(1-0.1,尺度参数)”。

④ 单击 OK 按钮运行, 则在数据编辑窗口中, 出现一系列变量名为 X 大于 x 的概率。见图 3-13。

⑤ 结论

x 应在大于 46.05 的范围内取值。

3.2.3.4 正态 (Normal) 分布

在随机变量的分布中, 正态分布是最最重要的一个分布。

1. 定义

(1) 一般正态分布定义

如果随机变量 X 的概率密度为:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

$-\infty < \mu < \infty$ 为常数, $\sigma > 0$ 为常数, 则称 X 服从正态分布。记作 $X \sim N(\mu, \sigma^2)$ 。

正态分布的分布函数为

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

(2) 标准正态分定义

如果随机变量 X 的概率密度为

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty$$

则称 X 服从标准正态分布。记作 $X \sim N(0, 1)$ 。

标准正态分布的分布函数为

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

当随机变量 $X \sim N(\mu, \sigma^2)$, 则通过数据转换 $U = \frac{x-\mu}{\sigma}$ 后, $U \sim N(0, 1)$ 。称 $U = \frac{x-\mu}{\sigma}$

为数据资料的标准化处理。

2. 由正态分布引出的抽样分布

统计量是样本的函数, 也是随机变量, 称其分布为抽样分布, 它对统计方法的应用起着举足轻重的作用。

由正态分布可以导出以下三个重要的连续型的抽样分布。

(1) 卡方分布

设随机变量 X_1, X_2, \dots, X_n 相互独立, 且均服从 $N(0, 1)$, 令

$$X = X_1^2 + X_2^2 + \dots + X_n^2$$

则 X 的分布称为具有自由度 n 的 χ^2 分布, 记作 $X \sim \chi^2(n)$ 。

χ^2 分布的概率密度为

$$f_X(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

其中 $\Gamma(a)$ 称为 Gamma 函数, 其定义为: $\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} dx, a > 0$ 。

当 Y 、 Z 相互独立且依次服从 $\chi^2(m)$ 及 $\chi^2(n)$, 那么, $Y+Z$ 服从 $\chi^2(m+n)$ 。当独立变量为多个时, 它同样成立。这说明卡方分布具有可加性。这一点对应用很重要。

此外, 费歇 (R.A.Fisher) 曾证明, 当 n 较大时, $\sqrt{2\chi^2(n)}$ 近似服从正态分布 $N(\sqrt{2n-1}, 1)$ 。

卡方分布的期望为: n , 卡方分布的方差为: $2n$ 。

(2) t (Student) 分布

设 $X \sim N(0, 1), Y \sim \chi^2(n)$, X 与 Y 独立, 则称随机变量 $T = \frac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的 t 分布, 又称学生氏 (Student) 分布, 记作 $T \sim t(n)$ 。

T 的概率密度为

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, -\infty < x < +\infty$$

t 分布的期望为: 0 (当 $n > 1$ 时), t 分布的方差为: $\frac{n}{n-2}$ (当 $n > 2$ 时)。

(3) F 分布

设 $X \sim \chi^2(m), Y \sim \chi^2(n)$, 且 X 与 Y 独立, 则随机变量 $\frac{X/m}{Y/n}$ 服从第一自由度 m , 第二自由度 n 的 F 分布, 记作 $F \sim F(m, n)$ 。

$F(m, n)$ 的密度函数为

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}}, x > 0$$

F 分布的期望为: $\frac{n}{n-2}$ ($n > 2$), F 分布的方差为: $\frac{2n^2(n+m-2)}{m(n-2)^2(n-4)}$ ($n > 4$)。

3. 正态分布的期望与方差

(1) 一般正态分布的期望与方差

一般正态分布的期望为: μ

一般正态分布的方差为： σ^2

(2) 标准正态分布的期望与方差

标准正态分布的期望为：0

标准正态分布的方差为：1

4. 应用背景

大量的实际经验与理论分析表明，测量误差及很多质量指标、人体的形态、机能、素质指标，如长度、强度、成绩等都是服从或近似服从正态分布。

5. SPSS 中与正态分布相关的函数

IDF.NORMAL(*prob, mean, stddev*)函数，它用来计算在给定的平均数 *mean* 和标准差 *stddev* 下，累积概率值为 *prob* 时的随机变量 *X* 的值。

CDF.NORMAL(*quant, mean, stddev*)函数是正态分布的累积函数，它用来计算在给定的平均数 *mean* 和标准差 *stddev* 下，小于 *quant* 值的累积概率值。

PDF.NORMAL(*quant, mean, stddev*)函数，它用来计算在给定的平均数 *mean* 和标准差 *stddev* 下，随机变量 *X* 在 *quant* 值时的概率密度值。

6. 数据资料的正态性检验

关于假设检验方面的基本知识详见第 6 章假设检验和第 7 章非参数假设检验。考虑到正态分布在实际中有重要的应用价值，故将对其的检验方法在此作一综述，读者也可先跳过本节的内容，待了解了第 6 和第 7 章的内容后，再回过头来阅读这部分内容。

例 3.13 用数据文件 data02-18.sav 中的 100 名健康成年女子血清蛋白含量的数据资料，在 SPSS 中对该样本数据资料做正态性检验。

在 SPSS 的很多过程里，都可以用来做数据资料正态性进行检验，归纳起来，有以下几种不同的方法，各种方法都有严格的使用条件的要求，切不可不顾条件拿来就用。

方法一 在 SPSS 中用 Nonparametric Tests 中 One-Sample Kolmogorov-Smirnov Test 过程对数据资料进行正态性检验

本程序中的渐近分布法要求样本含量大于等于 100，不满足要求时，可选用其 Exact 检验法和蒙特卡罗法。

其具体操作过程为：

① 在 SPSS 数据编辑窗口中，打开 data02-18.sav。

② 按 Analyze→Nonparametric Tests→1-sample K-S 顺序逐一单击鼠标键，展开 One-sample Kolmogorov-Smirnov Test 对话框（见图 3-7）。在左侧变量名框中，选择血清总蛋白，并将其移入到 Test Variable List 框中。

③ 在 Test Distribution 框中选中 Normal 复选项。

④ 单击 OK 按钮执行，在输出窗中出现计算结果，见表 3-4。

在表 3-4 中，因为 One-Sample Kolmogorov-Smirnov Test 表中最后一行的 Asymp.Sig.(2-tailed) 为 $0.842 > 0.05$ ，所以不拒绝总体服从正态分布的原假设。

方法二 用 Lilliefors (1967) 修正检验法进行数据资料的正态性检验

当样本含量小于 100，总体均数和标准差未知时，此时，用样本的均数和标准差作为总体均数和标准差的估计，再用柯尔莫哥洛夫检验统计量 D_n 计算，但这时统计量在原假设下的分布改变了，所以，不能再用柯尔莫哥洛夫检验时的临界值表，应改用 Lilliefors (1967) 修正的临界值表进行判断，即所谓的 Lilliefors (1967) 修正检验法。

本例可用本法来检验，在 SPSS 中具体操作步骤如下：

① 打开数据文件 “data02-18.sav”。

② 输入检验变量

按 Analyze → Descriptive Statistics → Explore，打开 Explore 对话框，见图 2-100。将血清蛋白变量移入到 Dependent List 框中。在 Display 中选择 Plots 选项。

③ 选择 Lilliefors 法数据资料正态性检验

按 Plots 按钮，展开 Plots 对话框，见图 2-105。选择 Normality plots with tests 选项，要求在输出窗输出图形与检验结果。按 Continue 返回 Explore 对话框。

按 OK 键，则在输出窗中得到输出结果。

这些输出结果有很多，可以用在本例检验正态分布这个特定场合的有用结果见表 3-5、图 3-14。

表 3-4 正态检验结果

One-Sample Kolmogorov-Smirnov Test	
	血清蛋白
N	100
	7.3600
	.39528
	.062
	.062
	.060
	.616
	.842

a. Test distribution is Normal.

表3-5 正态性检验结果

Tests of Normality						
	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
血清蛋白	0.62	100	.200*	.990	100	.699

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

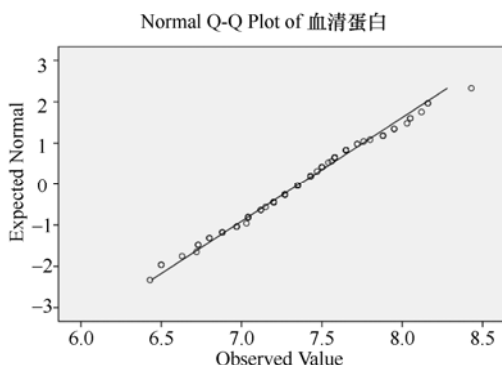


图 3-14 血清蛋白的正态 Q-Q 图

④ 结果与分析

从表 3-5 的正态检验表可见, Kolmogorov-Smirnov 检验的最大差值统计量为 0.62, 自由度为 100, 原假设成立的概率为大于 0.2。

Shapiro-Wilk 检验的统计量为 0.990, 自由度为 100, 原假设成立的概率为 0.699。

在图 3-14 中的 Q-Q 正态性检验中, 因为极大部分数据在其斜线上, 重合程度很高, 说明血清蛋白数据资料基本服从正态分布。

⑤ 结论

因为原假设成立的概率大于 0.05, 所以不拒绝血清蛋白数据资料服从正态分布的原假设。

方法三 用偏度和峰度指标对数据资料进行正态性检验

偏度和峰度是描述数据分布特征的统计量, 偏度可用来反映分布的偏斜方向和程度, 而峰度可用来反映分布的陡峭和平坦的程度, 加上非对称分布必向某侧偏斜之特点, 因而, 考察数据资料分布是否为正态分布时, 可用这两个指标来做比较性度量。

① 偏度和峰度的计算公式及判别标准

有关偏度和峰度的计算公式及判别标准见第 2 章 2.3.4 中的内容。

② 对样本含量的要求

样本含量应大于 200。有的学者认为, 计算峰度时, 样本含量应大于 1000。

③ SPSS 中具有计算偏度和峰度指标的过程汇总

在 SPSS 中, 具有计算偏度和峰度指标的子菜单见表 3-6。

表 3-6 在 SPSS 中有偏度和峰度指标的过程汇总

菜单	子菜单	程序	对话框	选择项	位于	备注
Analyze	Reports	OLAP Cubes	OLAP Cubes	Statistics	Statistics 变量列表	
		Case Summaries	Case Summaries			
		Reports Summaries in Rows	Reports: Summaries in Rows	Summary	Final Summary Lines	
		Reports Summaries in Columns	Reports: Summaries in Columns		Reports: Summary Lines for	
	Descriptive Statistics	Frequencies	Frequencies	Statistics	Distribution	在输出结果中
		Descriptives	Descriptives	Options		
		Explore	Explore	Statistics	选择 Descriptives	
	Compare means	means	means	Options	Statistics	单变量时, 将其复制一个变量

注意：由于本例的样本含量不足 200，故严格意义上来说不能用偏度和峰度指标来判定数据资料是否服从正态分布。

方法四 用卡方检验（Nonparametric Tests→Chi-square test）法对数据资料进行正态分布检验

本法的适用条件是数据资料为计数资料。

基本步骤如下：

① 首先将血清蛋白的数据资料整理成频数分布表，将计量资料变成计数资料。具体做法可参见第 2 章中介绍的计量资料频数分布表制作方法。

本例中根据计量资料频数分布表制作方法将 100 名健康成年女子血清蛋白数据制作成了频数分布表，见表 2-15。

为减少统计误差，需将组内频数不足 5 的组同相邻组合并，使合并后的组内频数大于 5，以便满足作卡方检验的要求。同理，将最后三组合并使合并后的组内频数等于 7 大于 5。这样一来，第一组的下限为 $-\infty$ ，近似上限为 6.8，而最后一组的下限为 8.0，上限为 $+\infty$ ，见表 3-7。为统计的方便及后续求理论概率的需要，我们取它为 8.6。这样，我们实际共分 8 组，各组的近似组上限分别为：6.8、7.0、7.2、7.4、7.6、7.8、8.0、8.6（应为 $+\infty$ ）。

表 3-7 合并数据后的频数分布表

组下限	Frequency	Percent	Valid Percent	Cumulative Percent
6.60	8	5.0	5.0	8.0
6.80	8	8.0	8.0	16.0
7.00	13	13.0	13.0	29.0
7.20	25	25.0	25.0	54.0
7.40	23	23.0	23.0	77.0
7.60	9	9.0	9.0	86.0
7.80	7	7.0	7.0	93.0
8.00	7	6.0	6.0	99.0
Total	100	100.0	100.0	

② 在 SPSS 中, 建立数据文件 data03-09.sav。其中有两个变量, 一个是组上限, 另一个是频数。将上述①中的各组的近似组上限值录入组上限中, 并将表 3-7 中的各组的合并后的频数顺序录入到频数中。

③ 如果血清蛋白含量这个变量服从正态分布, 则可用现有样本的均数及标准差作为其所隶属总体的均数及标准差的无偏估计, 因此, 在 SPSS 中, 选用第 1 章介绍的任一种描述统计的方法可求得样本的均数及标准差。本例样本的均数及标准差分别为: 7.36 和 0.3953。

④ 计算各组的理论期望频数 (或概率)

在 SPSS 中可用函数 CDF.NORMAL(*quant,mean,stddev*)先计算各组的累积频数 (或累积概率), 再用建立时间系列中计算差值的方法求各组理论期望频数 (或期望概率)。具体做法如下:

在 SPSS 的数据编辑窗口, 打开 data03-09.sav。

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。

在 Target Variable 框中, 输入目标变量名为各组理论累积频数。在 Numeric Expression 框中, 输入 “100*CDF.NORMAL(组上限, 7.36,0.3953)”。(如果输入的表达式中不乘 100, 则计算得到的是各组的累积概率。)单击 OK 按钮执行, 则在工作的数据文件中, 出现各组理论累积频数的新变量及其计算结果。

再按 Transform→Create Time Series 顺序, 打开 Create Time Series 对话框, 见图 3-15。

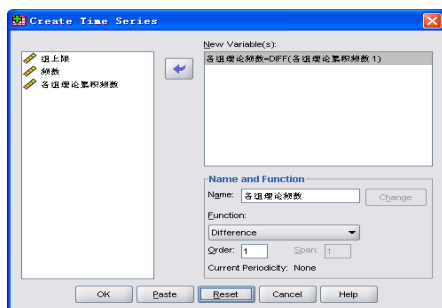


图 3-15 Create Time Series 主对话框

在左侧变量名源框中，选择各组理论累积频数，将移入到 New Variable(s)框中。

在 Name and Function 下的 Name 框中，修改变量名为：各组理论频数，单击 Change 按钮，则在 New Variable(s)框中出现对各组理论累积频数转换的方法及生成的新变量名称。

其他保持系统默认的选项（即在 Function 的下拉式选项确信选择的是 Difference 选项），单击 OK 按钮执行，在工作的数据文件中，出现各组理论频数的新变量及其计算结果。见图 3-16。

	组上限	频数	各组理论累积频数	各组理论频数
1	6.80	8.00	7.83	.
2	7.00	8.00	18.12	10.29
3	7.20	13.00	34.28	16.16
4	7.40	25.00	54.03	19.75
5	7.60	23.00	72.81	18.78
6	7.80	9.00	86.72	13.90
7	8.00	7.00	94.73	8.01
8	8.60	7.00	99.91	5.19

图 3-16 各组理论频数的计算结果

在图 3-16 中可见，在每组理论频数中第一组中理论频数是缺失值，它应等于同组的各组理论累积频数的值 7.83。可以用复制粘贴的方法人工将其补上。

⑤ 进行分布拟合优度检验即卡方检验。步骤如下：

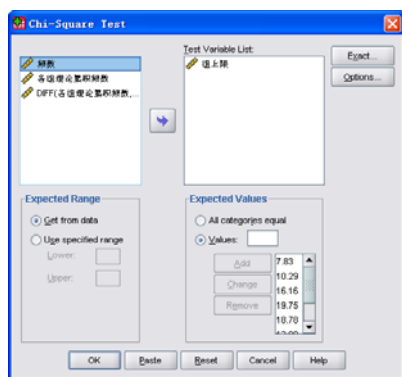


图 3-17 Chi-Square Test 对话框

用频数变量做加权处理，方法参见第 2 章例 2.21 中的做法。

按 Analyze→Nonparametric Tests→Chi-Square 顺序展开 Chi-Square Test 对话框，见图 3-17。

选择组上限变量，将其移入到 Test Variable List 框中。

在 Expected Values 栏中选择 Values 项，并在 Values 参数框中输入 7.83，按其下面的 Add 按钮，将 7.83 增加到期望值框中，重复这个过程，分别依次键入各组理论频数中的数据 10.29、16.16、19.75、18.78、13.90、8.01 直至 5.19。将它们全部有序地添加到期望值框中。

如果，数据文件中没有理论期望频数，而只有理论期望概率，则在 Values 中依次输入各组的理论概率值也是可以的，不影响最终的计算结果。

单击 OK 按钮，提交运算。结果如下：见表 3-8。因 $P=0.545>0.05$ ，故不拒绝血清蛋白含量这个变量服从正态分布的原假设。

表 3-8 血清蛋白含量正态分布检验结果

组上限				Test Statistics	
	Observed N	Expected N	Residual		组上限
6.80	8	7.8	.2	Chi-Square ^a	5.956
7.00	8	10.3	-2.3	df	7
7.20	13	16.2	-3.2	Asymp. Sig.	.545
7.40	25	19.8	5.2	a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 5.2.	
7.60	23	18.8	4.2		
7.80	9	13.9	-4.9		
8.00	7	8.0	-1.0		
8.60	7	5.2	1.8		
Total	100				

7. 正态分布的应用

(1) 确定参考值范围

医学上的参考值范围是指正常人体的解剖、生理、生化、免疫及组织代谢产物的含量等各种数据的波动范围。它的确切含义为从选择的参照总体上获得的所有检查结果，用统计的方法建立百分位数界限时所得到的区间称为**参考值范围**。习惯上是包含 95% 的参照总体的范围。

如果选用参照样本来确定参考值范围，则样本含量必须足够，一般应大于 120 以上。数据资料不呈正态分布时，可用百分位数来确定。

参考值范围对应正态分布的区间见表 3-9。

表 3-9 参考值范围所对应的正态分布区间

百分范围 (%)	单侧		双侧 (对称)	
	下限	上限	下限	上限
95	$\bar{X} - 1.65S$	$\bar{X} + 1.65S$	$\bar{X} - 1.96S$	$\bar{X} + 1.96S$
99	$\bar{X} - 2.33S$	$\bar{X} + 2.33S$	$\bar{X} - 2.58S$	$\bar{X} + 2.58S$

例 3.14 已知正常成年男子的红细胞数近似服从正态分布，测得某地正常成年男子的红细胞数的均数为 $4.78 \times 10^{12} / L$ ，标准差为 $0.38 \times 10^{12} / L$ ，估计该地成年男子红细胞数的 95% 的参考值范围。

【题析】 本例已知数据资料近似服从正态分布，故可利用正态分布的理论来处理。因本例中红细胞数过多或过少均属异常，故采用双参考值范围。

在 SPSS 中，当数据资料已经被计算成本例的统计指标时，可用如下的方式进行计算：

① 在 SPSS 数据编辑器中，建立数据文件

在数据编辑器中建立如图 3-18 所示的数据文件，这两个变量均为标准数值型尺度测度变量。

	平均数	标准差
1	4.78	.38

图 3-18 数据输入样式

② 调用 $IDF.NORMAL(prob, mean, stddev)$ 函数计算上、下限值。

按 Transform → Compute 顺序展开 Compute 对话框 (见图 2-53)。在 Target Variable 下框中输入需要建立的一个新变量名 下限, 将鼠标光标移到 Numeric Expression 下框并单击, 在 Function group 下框中单击 All, 在 Functions and Special Variables 下框中选择 $IDF.NORMAL(prob, mean, stddev)$ 函数双击, 则在 Numeric Expression 下框中出现 $IDF.NORMAL(prob, mean, stddev)$ 函数。选中该函数的第一个问号, 输入 0.025, 选中该函数的第二个问号, 单击变量名列表中的平均数, 按中间的右移按钮, 则平均数覆盖在函数的第二个问号上, 成为函数的第二个参数, 同样, 选中该函数的第三个问号, 单击变量名列表中的标准差, 按中间的右移按钮, 则标准差覆盖在函数的第三个问号上, 成为函数的第三个参数, 这样, 则完成对函数参数的赋值。单击 OK 按钮, 则在数据窗口工作的数据文件中出现下限新变量及其计算值。

重复上述这个过程, 将 Target Variable 下框中的变量名换为 上限, 第一个参数换为 0.975 即可。则在数据窗工作的数据文件中出现上限新变量及其计算值, 见图 3-19。

	平均数	标准差	下限	上限
1	4.78	.38	4.04	5.52

图 3-19 计算结果

③ 结果

该地成年男子红细胞数的 95% 的参考值范围为 $4.04 \times 10^{12}/L$ 至 $5.52 \times 10^{12}/L$ 。

当有原始数据时, 可直接用描述统计中的探究性分析方法, 一方面对原始数据作正态性检验, 如果接受正态分布的原假设, 则选用描述统计中的频数过程, 在 Statistics 选择项中, 选择 Percentile(s) 选项, 依次输入 2.5、97.5 (或 5、或 95), 即可得到 95% 的参考值范围的粗略估计值。

(2) 估计总体均数 95% 的置信区间

在只有统计量: 平均数、标准差、样本含量、标准误差的情况下, 当样本数据资料服从正态分布时, 可仿例 3.14 中的做法, 只需将标准差变量用标准误差替代 (标准误差 = 标准差 / 样本含量的算术平方根), 即可得到总体均数 95% 的置信区间。当已知的是原始数据时, 在 SPSS 中, 求总体均数 95% 的置信区间的做法也有几种, 最常用的可仿第 2

表 3-10 计算总体均数 95% 的置信区间

Descriptives			Statistic	Std. Error
血清总蛋白	Mean		7.3600	.03953
	95% Confidence Interval for Mean	Lower Bound	7.2816	
		Upper Bound	7.4384	
	5% Trimmed Mean		7.3589	
	Median		7.3500	
	Variance		.156	
	Std. Deviation		.39528	
	Minimum		6.43	
	Maximum		8.43	
	Range		2.00	
	Interquartile Range		.46	
	Skewness		.077	.241
	Kurtosis		.052	.478

章例 2.47 中的操作过程, 在其输出结果中, 可得到总体均数的 95% 的置信区间, 即用描述统计中的探究性分析方法可以直接求原始数据的总体均数 95% 的置信区间。

例 3.15 在例 3.13 中, 已经对 100 名健康成年女子血清蛋白含量的数据资料, 在 SPSS 中进行了正态性检验, 结果认为血清蛋白含量随机变量服从正态分布, 试求其总体均数 95% 的置信区间 (其数据资料存放在数据文件 data02-18.sav 中)。

在 SPSS 中, 打开数据文件 data02-18.sav。按第 2 章例 2.47 操作过程, 选择血清总蛋白将其移入到 Dependent List 框中, 其他操作过程同例 2.47, 可在输出窗口中得到表 3-10。

在表 3-10 中可见, 成年女子血清蛋白含量总体均数 95% 的置信区间的下限值为: 7.2816, 上限值为: 7.4384。

(3) 已知概率值求随机变量 X 值

此类问题有共性, 都是知道各个等级出现的概率, 反求 X 。所以, 可以用 $IDF.NORMAL(prob, mean, stddev)$ 来进行计算 X 的值, 需要注意的是参数 $prob$ 指的是累积概率。

例 3.16 在将萨摩梅子的重量按正态分布分等时, 其中 20% 为小的, 55% 为中等, 15% 为大的, 另外 10% 为很大的。如果全部萨摩梅子的平均重量为 4.83 盎司, 标准差为 1.20 盎司, 试求中号萨摩梅子重量的上下限。

在 SPSS 中, 处理本类问题的操作步骤如下:

- ① 在 SPSS 中, 建立数据文件, 见 data03-10.sav。
- ② 按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框。在 Target Variable 框中, 输入目标变量名为 *中号的下限*。
- ③ 在 Numeric Expression 框中, 输入 “ $IDF.NORMAL(\text{小}, \text{平均数}, \text{标准差})$ ”。
- ④ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现一系列变量名为中号的下限的变量名及计算值。见图 3-20。

	平均数	标准差	小	中	大	很大	中号的下限	中号的上限
1	4.83	1.20	0.20	0.55	0.15	0.10	3.82	5.64

图 3-20 中号的下限和上限

重复 2、3 过程, 在 Target Variable 框中, 输入目标变量名为 *中号的上限*。在 Numeric Expression 框中, 输入 “ $IDF.NORMAL(\text{小}+\text{中}, \text{平均数}, \text{标准差})$ ”。单击 OK 按钮运行, 则在数据编辑窗口中, 出现一系列变量名为 *中号的上限* 的变量名及计算值。见图 3-20。

例 3.17 某车间有 300 台独立工作的车床, 在生产时间内由于需要检修, 调换工具等常需要停车, 假设其开工率为 0.7, 每台车床工作时需电力 1 千瓦, 问应至少给该车间供多少瓦电力时, 才能有 99.9% 的可能性保证该车间不会因供电不足影响生产。

【题析】根据列维-林德贝格 (Levy-lindeberg) 中心极限定理, 可以得到棣莫佛-拉普拉斯 (De Moivre-Laplace) 推论:

设 μ_k 表示 n 次独立重复试验中事件 A 发生的次数, p 是事件 A 在每次试验中发生的概率, $0 < p < 1$, 则对任意的 $x \in (-\infty, +\infty)$, 有

$$\lim_{n \rightarrow \infty} P\left\{\frac{\mu_n - np}{\sqrt{npq}} \leq x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$$

其中 $q = 1 - p$ 。

据此, 可以知道, 在大样本情况下, n 次独立重复试验中事件 A 发生的次数近似服从正态分布。因而, 可以用二项分布的期望 np 代替正态分布的平均数, 用二项分布的标准差 \sqrt{npq} 代替正态分布的标准差。

在本例中, $\bar{X} = np = 300 \times 0.7 = 210$, $S = \sqrt{npq} = \sqrt{300 \times 0.7 \times 0.3} = 7.9373$ 。每供应 1 千瓦的电力相当于一台车床在工作。

因此, 本例可转化为: 需要多少台车床同时工作时才有 99.9% 的可能性保证该车间不影响生产。

在 SPSS 中, 处理本类问题的操作步骤如下:

- ① 在 SPSS 中, 建立数据文件, 见 data03-11.sav。
- ② 按 Transform → Compute Variable 顺序, 打开 Compute Variable 对话框。在 Target Variable 框中, 输入目标变量名为至少需要的电力量。
- ③ 在 Numeric Expression 框中, 输入 “IDF.NORMAL (开工率, 车床*开工率, sqrt (车床*开工率* (1-开工率)))”。
- ④ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现一系列变量名为至少需要的电力量的变量名及计算值。见图 3-21。

	车床	开工率	至少需要的电力量
1	300.00	0.70	214.16

图 3-21 计算结果

因此, 至少给该车间供 215 千瓦的电力时, 才能有 99.9% 的可能性保证该车间不会因供电不足影响生产。

(4) 已知随机变量 X 值求概率值

同 (3) 一样, 此类问题也都有共性, 都是在事先知道随机变量 X 的观察值 x 的条件下, 来求事件大于 x 或小于 x 出现的概率。所以, 可以用函数 CDF.NORMAL (quant, mean, stddev) 来计算随机变量 X 的观察值小于等于给定的 quant 的概率。

例 3.18 有三个学生参加不同测验，测验成绩都服从正态分布。A 得 72 分，B 得 85 分，C 得 17 分。与 A 参加测验的全部学生平均成绩为 85 分、标准差为 7 分，与 B 参加测验的全部学生平均成绩为 90 分、标准差为 3 分，与 C 参加测验的全部学生平均成绩为 25 分、标准差为 7 分。按照这些平均数和标准差，你可否判定这三个学生的成绩优异的次序，并说明理由。

在 SPSS 中，处理本类问题的操作步骤如下：

- ① 在 SPSS 中，建立数据文件，见 data03-12.sav。
- ② 按 Transform→Compute Variable 顺序，打开 Compute Variable 对话框。在 Target Variable 框中，输入目标变量名为 *低于本得分的概率*。
- ③ 在 Numeric Expression 框中，输入“CDF.NORMAL (得分，平均数，标准差)”。
- ④ 单击 OK 按钮运行，则在数据编辑窗口工作的数据文件中，出现一系列变量名为 *低于本得分的概率* 的变量名及计算值。见图 3-22。

	平均数	标准差	得分	低于本得分的概率	该得分对应的 u 值
1	85.00	7.00	72.00	0.03	-1.86
2	90.00	3.00	85.00	0.05	-1.67
3	25.00	7.00	17.00	0.13	-1.14

图 3-22 小于该得分的人数百分比和该得分对应的 u 值

从图 3-22 可知，低于 A 得分的概率为 0.03，说明 A 在参加同类测验的全部学生中，成绩从低到高排在 3% 的位置，同理，B 在参加同类测验的全部学生中，成绩从低到高排在 5% 的位置，C 在参加同类测验的全部学生中，成绩从低到高排在 13% 的位置，因此，据此可以判定三个学生的优异的次序，C 最好、B 次之、A 最差。

本例还可以用三个学生在各自正态分布总体中的位置值，来加以判定各自的优劣。

步骤同上。不同之处只需在 Target Variable 框中，输入目标变量名为 *该得分对应的 u 值*。在 Numeric Expression 框中，输入“(得分-平均数)/标准差”。则执行完上述步骤后，在工作数据文件中得到如图 3-22 所示的计算结果。

由于 C 的正态分布位置值 (-1.14) > B 的正态分布位置值 (-1.67) > A 的正态分布位置值 (-1.86)，所以，由此可以判定三个学生的优异的次序，C 最好、B 次之、A 最差。

例 3.19 按美国经验死亡表，年龄为 41 岁的人在一年内死去的概率是 0.01。一家保险公司对于这一年龄的人有 50 000 张有效保险票，试用正态分布的理论估计该公司在一年内对于这群人中有 525 个以上要求死亡赔款的概率。

【题析】 根据棣莫佛-拉普拉斯定理，在大样本情况下， n 次独立重复试验中事件 A 发生的次数近似服从正态分布。

本例, $\bar{X} = np = 50000 \times 0.01 = 500$, $S = \sqrt{npq} = \sqrt{50000 \times 0.01 \times 0.99} = 6.7082$ 。

因此, 本例可以转化为在均数为 500, 标准差为 6.7082 的正态分布中, 求大于 525 的概率。

在 SPSS 中, 进行本例操作的具体步骤如下:

① 在 SPSS 中, 建立数据文件, 见 data03-13.sav。

② 按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框。在 Target Variable 框中, 输入目标变量名为 *超过 525 人的概率*。

③ 在 Numeric Expression 框中, 输入 “1-CDF.NORMAL (要求死亡赔款的人数, 死亡概率*样本含量, sqrt (样本含量*死亡概率*(1-死亡概率)))”, 其中 Sqrt (numexpr) 是求算术平方根函数。

④ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现一列变量名为低于本得分的概率的变量名及计算值。见图 3-23。

	死亡概率	样本含量	要求死亡赔款的人数	超过525人的概率
1	0.01	50000.00	525.00	0.13

图 3-23 大于 525 人以上要求死亡赔款的概率

⑤ 结果

该公司在一年内对于这群人中有 525 个以上要求死亡赔款的概率为 0.13。

第4章 参数估计

在第3章中,我们讨论了离散型和连续型变量最常见的几种分布类型及其应用。在实际的研究中,总体分布的参数可能是已知的,也可能是未知的。这就需要我们通过对所收集到的样本进行研究,根据样本提供的信息,对其总体的分布或分布参数做出统计推断。推断统计是统计学的重要内容之一。它主要包括参数估计与假设检验两大部分。本章主要讨论参数估计。

一般来说,当研究的总体分布类型已知,其参数中存在未知参数,而研究的目的又需要完全了解其分布时,需用到参数估计方法。例如,已知 $X \sim B(n, p)$,即随机变量 X 服从二项分布,但 p 未知,则 X 的分布仍不算是完全已知的,此时,如要计算出 X 的概率,需要完全知道它的分布,因此,首先必须对未知参数 p 进行估计。此外,有时总体的分布类型也可能未知,我们只想知道它的数学期望、方差等的数字特征,也需要用到参数估计的方法。总的说来,对分布中的未知参数或未知分布总体中的某些数字特征进行估计的问题称为参数估计问题,简称参数估计。

通常我们用 θ 表示总体的一个未知参数,用 θ_l 表示总体的第 l 个未知参数。为实现对总体未知参数 θ (或 θ_l)进行估计,一个很重要的方法就是从所要研究的总体中用简单随机抽样的方法抽取一个有足够样本含量 n 的样本 X_1, X_2, \dots, X_n ,这样所得到的样本同总体 X 有相同的概率分布,此时,可用样本所提供的信息,对总体 X 的属性进行估计、推断。

在第3章中,所介绍的频数分布表的制作方法,就是当总体分布类型未知时,用大样本数据对总体分布密度的一种近似求法。

通常有两种基本的方法可用来估计参数的大小,一种是点估计,另一种是区间估计。

4.1 参数的点估计

设 X_1, X_2, \dots, X_n 是来自总体 X 的样本, X 的分布类型已知, θ 是总体的未知参数,则点估计是用一统计量 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 作为参数 θ 的估计。称估计用的统计量 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 为 θ 的估计量。实测一个样本 x_1, x_2, \dots, x_n ,将样本观测值代入估计量可得到 θ 的估计值 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 。该估计量的值可作为参数的一个近似值。

这种用一个统计量来估计总体的某个未知参数的方法称为**参数的点估计方法**。它还可以推广到参数为多个的情形。

设总体 X 的分布类型已知, 但 X 中含有 l 个不同的未知参数: $\theta_1, \theta_2, \dots, \theta_l$, 则来自总体 X 的样本 X_1, X_2, \dots, X_n 建立的 l 个不含任何未知参数的统计量 $\hat{\theta}_i(X_1, X_2, \dots, X_n) (i=1, 2, \dots, l)$ 称为未知参数 $\theta_i (i=1, 2, \dots, l)$ 的估计量。实测一个样本 x_1, x_2, \dots, x_n , 将样本观测值代入估计量可得到 θ_i 的估计值 $\hat{\theta}_i(x_1, x_2, \dots, x_n)$ 。它代表 l 维参数空间 (指参数所有可能的取值范围, 记作 Θ) 中的一个点。

因此, 点估计就是给出参数空间中的一个点。一个参数的点估计是数轴上的一个点, 两个参数的点估计是平面上的一个点, l 个参数的点估计是 l 维参数空间中的一个点。

例如, 当分布类型已知, $X \sim N(\mu, \sigma^2)$, 而正态总体的参数 μ 和 σ^2 未知时, 参数的点估计中, 常用样本均值 \bar{X} 作为总体均值 $E(X) = \mu$ 的估计量, 用样本的方差 S^2 作为总体

方差 $D(X) = \sigma^2$ 的估计量。实测一个样本 x_1, x_2, \dots, x_n , 得到 μ 的估计值为 $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$, σ^2 的

估计值为 $S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ 。

在实际的估计中, 对同一个参数 θ , 有许多不同的估计量, 即有许多不同的估计方法。但最常用的方法为**矩估计法**和**极大似然估计法**两种。

4.1.1 参数的矩估计法

矩估计法是参数估计中一种被广泛使用的方法, 它由英国统计学家 K. Pearson 于 1894 年提出。

随机变量的矩是分布的重要数字特征。因此, 矩估计法也叫数字特征法。在概率论中, 随机变量的矩的定义:

X 的 k 阶原点矩定义为: $a_k = \int_{-\infty}^{+\infty} x^k dF(x)$, 它是随机变量 X 取值的 k 次方的加权平均。

X 的 k 阶中心矩定义为: $\beta_k = \int_{-\infty}^{+\infty} (x - E(X))^k dF(x)$, 它是对 X 进行“中心化”以后所得差值的 k 次方的加权平均。

在上面的定义中, $F(X)$ 为 X 的分布函数。

当 X 为连续型随机变量时, 其密度函数为 $f(x)$, 则 $a_k = \int_{-\infty}^{+\infty} x^k f(x) dx$; 当 X 为离散型随机变量时, 其概率函数为 $P(X = x_i) = p_i (i = 1, 2, \dots)$, 则 $a_k = \sum_{i=1}^{+\infty} x_i^k p_i$ 。当这些积分或级数绝对收敛时, a_k 存在。对于中心矩也有类似的定义。

实际上, 把“样本均值 \bar{X} 作为总体均值 $E(X) = \mu$ 的估计量”这一做法进行推广就是矩估计法。这是因为, 从 k 阶原点矩定义中我们知道, 当 $k=1$ 时, 样本均值就是一阶样本原点矩, 而总体均值就是一阶总体原点矩, 将样本均值 \bar{X} 作为 $E(X)$ 的估计量的做法就是把一阶样本原点矩作为一阶总体原点矩的估计量。推广之就是, 把二阶样本原点矩作为二阶总体原点矩的估计量, 把 l 阶样本原点矩作为 l 阶总体原点矩的估计量。这些估计方法就是矩估计法。

求矩估计的具体做法为:

设总体 X 的分布类型已知, 其参数 θ 为 l 维, $\theta = (\theta_1, \theta_2, \dots, \theta_l)^T$, 即总体含有 l 个未知参数, 从总体中抽取一个样本 X_1, X_2, \dots, X_n , 当总体的一阶到 l 阶原点矩存在的前提下, a_1, a_2, \dots, a_l 一般都可表示为 θ 的函数: $a_k = a_k(\theta_1, \theta_2, \dots, \theta_l) (k = 1, 2, \dots, l)$ 。用样本矩代替总体矩, 可得到下面关于 $\theta = (\theta_1, \theta_2, \dots, \theta_l)^T$ 的方程组

$$\begin{cases} a_1(\theta_1, \theta_2, \dots, \theta_l) = E(X) = \frac{1}{n} \sum_{i=1}^n X_i \\ a_2(\theta_1, \theta_2, \dots, \theta_l) = E(X^2) = \frac{1}{n} \sum_{i=1}^n X_i^2 \\ \vdots \\ a_l(\theta_1, \theta_2, \dots, \theta_l) = E(X^l) = \frac{1}{n} \sum_{i=1}^n X_i^l \end{cases}$$

解上面的方程组可得 $\theta_1, \theta_2, \dots, \theta_l$ 的估计量为

$$\begin{cases} \hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \dots, X_n) \\ \hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \dots, X_n) \\ \vdots \\ \hat{\theta}_l = \hat{\theta}_l(X_1, X_2, \dots, X_n) \end{cases}$$

由上述方法所确定的参数的估计量称为矩估计量。

例 4.1 1900 与 1969 年间, 36 次台风侵入 Appalachians 地区时 24 小时的最大降雨量 (cm) 的数据资料见表 4-1。试对最大降雨量拟合 Gamma 分布。

表 4-1 最大降雨量 (cm)

年份	最大降雨量	年份	最大降雨量	年份	最大降雨量	年份	最大降雨量
1969	31.00	1952	4.95	1932	4.75	1916	7.43
1968	2.82	1949	5.64	1932	6.85	1915	5.00
1965	3.92	1945	5.51	1929	6.25	1915	4.58
1960	4.02	1942	13.40	1928	3.42	1912	4.46
1959	9.50	1940	9.72	1928	11.80	1906	8.00
1957	4.50	1939	6.47	1923	0.80	1902	3.73
1955	11.40	1938	10.16	1923	3.69	1901	3.50
1954	10.71	1934	4.21	1920	3.10	1900	6.20
1954	6.31	1933	11.60	1916	22.22	1900	0.67

解：设所求的 Gamma 分布的密度函数为

$$f(x, \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

要求最大降雨量拟合 Gamma 分布，实际上就是要求 Gamma 分布的密度函数中的 α 和 β 两个待定参数的值。

由 X 的 k 阶原点矩定义可得

$$\begin{aligned} a_1 &= \int_{-\infty}^{+\infty} x f(x, \alpha, \beta) dx = \int_0^{+\infty} \frac{\beta^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\beta x} dx = \frac{1}{\Gamma(\alpha)\beta} \int_0^{+\infty} t^\alpha e^{-t} dt \\ &= \frac{1}{\Gamma(\alpha)\beta} \Gamma(\alpha+1) = \frac{\alpha}{\beta} \end{aligned}$$

$$\begin{aligned} a_2 &= \int_{-\infty}^{+\infty} x^2 f(x, \alpha, \beta) dx = \int_0^{+\infty} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha+1} e^{-\beta x} dx = \frac{1}{\Gamma(\alpha)\beta^2} \int_0^{+\infty} t^{\alpha+1} e^{-t} dt \\ &= \frac{1}{\Gamma(\alpha)\beta^2} \Gamma(\alpha+2) = \frac{(\alpha+1)\alpha}{\beta^2} \end{aligned}$$

所以

$$\begin{cases} a_1 = \frac{\alpha}{\beta} = E(X) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \\ a_2 = \frac{(\alpha+1)\alpha}{\beta^2} = E(X^2) = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases}$$

解之得

$$\begin{cases} \hat{\alpha} = \frac{n\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{n\bar{X}^2}{(n-1)S^2} \\ \hat{\beta} = \frac{n\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{n\bar{X}}{(n-1)S^2} \end{cases}$$

因此，在 SPSS 中，要求解 Gamma 分布的密度函数中的 α 和 β 两个待定参数的值，根据公式就要先求出样本的均值和方差，这可用如下的步骤进行：

① 将表 4-1 中的数据在 SPSS 的数据编辑窗口中制成数据文件，见 data04-01.sav。

② 按 Analyze→Reports→Case Summaries 顺序，打开 Case Summaries 对话框（见图 2-51）。

在左边的变量名源框中，选中降雨量变量，通过中间的右移箭头将它移到 Variables 框中。

③ 单击 Statistics 按钮，打开 Statistics 对话框（见图 2-52）。

在 Statistics 对话框中选择 Number of Cases、Mean、Variance 统计量，并将其移入到 Cell Statistics 下框中。

单击 Continue 按钮，返回 Case Summaries 对话框。

④ 单击 OK 按钮，在输出窗口中得到样本的平均数和方差的计算结果，见表 4-2。

⑤ 求估计值并写出拟合密度函数

将表 4-2 中的计算结果代入到方程组的解中可得

$$\hat{\alpha} = 1.634, \quad \hat{\beta} = 0.224$$

所以，所求的拟合密度函数为

$$f(x) = \frac{0.224^{1.634}}{\Gamma(1.634)} x^{0.634} e^{-0.224x} \quad (x > 0)$$

表 4-2 平均数、方差的计算结果

Case Summaries

降雨量

N	Mean	Variance
36	7.2858	33.421

例 4.2 某自动化车床加工的某种零件的长度 X 是个随机变量， $X \sim N(\mu, \sigma^2)$ ，但参数 μ, σ^2 未知，现随机抽取该车床生产的零件 16 个，测得其长度资料如下（单位：cm）：
12.15, 12.12, 12.21, 12.01, 12.08, 12.09, 12.03, 12.01, 11.06, 12.13, 12.07, 12.11, 12.08, 12.01, 12.03, 12.06

试求总体 μ 和 σ^2 的估计值。

解：由矩估计法得

$$\begin{cases} a_1(\mu, \sigma^2) = E(X) = \frac{1}{n} \sum_{i=1}^n X_i \\ a_2(\mu, \sigma^2) = E(X^2) = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases}$$

即

$$\begin{cases} \mu = \bar{X} \\ \mu^2 + \sigma^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases}$$

所以，解之得

$$\begin{cases} \mu = \bar{X} \\ \sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S^2 \end{cases}$$

于是， μ 和 σ^2 的矩估计量为

$$\begin{cases} \hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S^2 \end{cases}$$

仿例 4.1 的做法，将题中数据做成数据文件 data04-02.sav。打开 data04-02.sav。按

表 4-3 平均数、方差计算结果

Case Summaries		
零件长度		
N	Mean	Variance
16	12.0156	.068

Analyze→Reports→Case Summaries 顺序，打开 Case Summaries 对话框（见图 2-51）。在左边的变量名源框中，选中零件长度变量，通过中间的右移箭头将它移到 Variables 框中。打开 Statistics 对话框（见图 2-52）。在 Statistics 对话框中选择 Number of Cases、Mean、Variance 统计量，并将其移入到 Cell Statistics 下框中。单击 Continue

按钮返回对话框。在对话框中，单击 OK 按钮，在输出窗口中得到样本的平均数和方差的计算结果，见表 4-3。

将表 4-3 中的计算结果代入到方程组的解中可得

$$\hat{\mu} = 12.0156, \quad \hat{\sigma}^2 = 0.06375$$

4.1.2 参数的极大似然估计法

用点估计来寻求估计量的另一种常用方法是极大似然估计法。它最早由高斯 (C.F.Gauss) 提出, 后由英国统计学家费歇 (R.A.Fisher) 证明了这个方法的一些性质, 并给出了“极大似然法”这个名称。应用表明, 这种方法有很多优良之处, 比如, 它改进了矩法和最小二乘法使得到的估计值的精度更高, 是一种非常重要的点估计方法。应用此法的前提条件是分布类型必须已知。

设总体 X 的 θ 为未知参数, 概率密度为 $f(x; \theta)$, X_1, X_2, \dots, X_n 为 X 的一个样本, 则称样本的联合概率密度为参数 θ 的似然函数, 记作 $L(\theta)$, 即

$$L(\theta) = L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

如果存在 $\hat{\theta}(x_1, x_2, \dots, x_n)$, 使其在 $\theta = \hat{\theta}$ 时

$$L(\hat{\theta}) = L(x_1, x_2, \dots, x_n; \hat{\theta}) = \max |L(x_1, x_2, \dots, x_n; \theta)| = \max |L(\theta)|$$

则称 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 为 θ 的极大似然估计量。

如果总体概率密度中有 l 个待估参数: $\theta_1, \theta_2, \dots, \theta_l$, 则参数 $\theta_1, \theta_2, \dots, \theta_l$ 的似然函数为:

$$L(\theta_1, \theta_2, \dots, \theta_l) = L(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_l) = \prod_{i=1}^n f(x_i; \theta_1, \theta_2, \dots, \theta_l),$$

如果存在 $\hat{\theta}_i(x_1, x_2, \dots, x_n) (i=1, 2, \dots, l)$, 使其在 $\theta_i = \hat{\theta}_i (i=1, 2, \dots, l)$ 时

$$\begin{aligned} L(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_l) &= L(x_1, x_2, \dots, x_n; \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_l) \\ &= \max |L(x_1, x_2, \dots, x_n; \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_l)| \\ &= \max |L(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_l)| \end{aligned}$$

则称 $\hat{\theta}_i = \hat{\theta}_i(X_1, X_2, \dots, X_n)$ 为 θ 的极大似然估计量。

极大似然估计的具体求法如下:

(1) 在参数为 θ 时, 对似然函数取对数得到

$$\ln L(\theta) = \sum_{i=1}^n \ln f(x_i; \theta)$$

对数函数为增函数, $L(\theta)$ 与 $\ln L(\theta)$ 在相同点达到最大, 如果一元函数极值存在, 则可得到似然方程

$$\frac{d \ln L(\theta)}{d \theta} = 0$$

求解该方程得到 θ 。即得 θ 的极大自然估计量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 。

(2) 在待估参数有 l 个时, 对似然函数取对数得到

$$\ln L(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_l) = \sum_{i=1}^n \ln f(x_i; \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_l)$$

同样地, 如果多元函数的极值存在, 则可得到似然方程组

$$\begin{cases} \frac{\partial \ln L(\theta_1, \theta_2, \dots, \theta_l)}{\partial \theta_1} = 0 \\ \frac{\partial \ln L(\theta_1, \theta_2, \dots, \theta_l)}{\partial \theta_2} = 0 \\ \vdots \\ \frac{\partial \ln L(\theta_1, \theta_2, \dots, \theta_l)}{\partial \theta_l} = 0 \end{cases}$$

由上面的似然方程组可解得 $\theta_1, \theta_2, \dots, \theta_l$, 从而得到 $\theta_1, \theta_2, \dots, \theta_l$ 的极大似然估计量为 $\hat{\theta}_i = \hat{\theta}_i(X_1, X_2, \dots, X_n) (i=1, 2, \dots, l)$ 。

例 4.3 已知总体 X 服从正态分布

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < +\infty)$$

求参数 μ 和 σ^2 的极大似然估计 $\hat{\mu}$ 和 $\hat{\sigma}^2$ 。

解: 由于总体 X 服从正态分布 $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < +\infty)$

所以, 参数 μ 和 σ^2 的似然函数为

$$L(\mu, \sigma^2) = \prod_{i=1}^n f(x_i; \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}}$$

上式两边同时取自然对数得

$$\begin{aligned} L(\mu, \sigma^2) &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \\ &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 - \frac{n}{2\sigma^2} (\bar{x} - \mu)^2 \end{aligned}$$

所以似然方程组为

$$\begin{cases} \frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{n}{\sigma^2}(\bar{x} - \mu) = 0 \\ \frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n}{2\sigma^4}(\bar{x} - \mu)^2 = 0 \end{cases}$$

解之得

$$\begin{cases} \mu = \bar{x} \\ \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{cases}$$

所以, 所求参数 μ 和 σ^2 的极大似然估计量为

$$\begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \end{cases}$$

它同矩估计的结果相一致。

例 4.4 把 500 条做记号的鱼放在一个鱼池中, 待这些鱼与鱼池中原来的鱼充分混和后, 再从鱼池中捕起 1000 条鱼, 发现其中 100 条有记号, 试估计鱼池中原有多少条鱼。

设鱼池中加入 500 条有记号的鱼后共有 N 条鱼, 则根据题意鱼池中原有 $N-500$ 条鱼, 并设从鱼池中捕起的 1000 条鱼中有记号的为 X 条, 没有记号的为 $1000-X$ 条。

显然 X 为随机变量, 它的密度函数为

$$P(X=k) = \frac{\binom{500}{k} \binom{N-500}{1000-k}}{\binom{N}{1000}}$$

故似然函数

$$L(N) = \frac{\binom{500}{k} \binom{N-500}{1000-k}}{\binom{N}{1000}}$$

由于 $L(N)$ 的定义域不是区间, 所以不能借助于求导来推算。因此, 在这种情况下, 可以通过似然函数的比值来辅助求解。

由于

$$\frac{L(N)}{L(N-1)} = \frac{\binom{500}{k} \binom{N-500}{1000-k}}{\binom{N}{1000}} \bigg/ \frac{\binom{500}{k} \binom{N-1-500}{1000-k}}{\binom{N-1}{1000}} = \frac{(N-1000)(N-500)}{[(N-500)(1000-k)]N}$$

当 $\frac{L(N)}{L(N-1)} > 1$ 时, $kN < 500000$, 当 $\frac{L(N)}{L(N-1)} < 1$ 时, $kN > 500000$, 所以, 当 $N < \frac{500000}{k}$ 时, $L(N-1) < L(N)$, 当 $N > \frac{500000}{k}$ 时, $L(N-1) > L(N)$, 故, 当 N 为不超过 $\frac{500000}{k}$ 的最大正整数时, 即 $N = \left\lfloor \frac{500000}{k} \right\rfloor$ 时, $L(N)$ 最大, 即 N 的最大似然估计值为 $\hat{N} = \left\lfloor \frac{500000}{k} \right\rfloor$ 。

又题中 $k=100$, 故 $\hat{N}=5000$, 所以鱼池中原有 4500 条鱼。

4.1.3 估计量的评选标准

总体的某个参数的估计量往往不是唯一的, 如期望 μ 的估计量, 可以选择 $\mu_1 = \bar{X}$, 也可以选取 $\mu_2 = X_1$ 。同样, 对方差 σ^2 的估计量, 在用矩估计和极大似然估计时, 都得到 $\hat{\sigma}_1^2 = \frac{n-1}{n} S^2$, 而平时用得最多的却是 $\hat{\sigma}_2^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = S^2$ 。

这样的例子有很多, 在有多个样本的函数都可以作为一个未知参数的估计量时, 究竟使用哪个估计量更好呢?

衡量估计量好坏的标准通常有三个, 一是无偏性, 二是有效性, 三是一致性。

设 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 是未知参数 θ 的估计量, 如果满足 $E(\hat{\theta}) = \theta$, 则称 $\hat{\theta}$ 具有无偏性, 亦称 $\hat{\theta}$ 为 θ 的无偏估计量。

这个标准指出, 对于无偏估计量 $\hat{\theta}$ 应满足无系统偏差的条件。由于 $\hat{\theta}$ 是个随机变量, 所以, 它取值应围绕参数的真值 θ 周围波动, 即 $\hat{\theta}$ 平均起来应与 θ 值相同, 这是无偏性的要求。当估计量不是无偏估计时, 则称它为有偏估计量。

总体未知参数 θ 的无偏估计量有时可以有很多, 例如, 设 X_1, X_2 是从总体 $N(\mu, 1)$ 中抽取的容量为 2 的样本, 则 $\hat{\mu}_1 = \frac{2}{3}X_1 + \frac{1}{3}X_2$ 、 $\hat{\mu}_2 = \frac{3}{4}X_1 + \frac{1}{4}X_2$ 、 $\hat{\mu}_3 = \frac{1}{2}X_1 + \frac{1}{2}X_2$, 由于 $E(\hat{\mu}_1) = E(\hat{\mu}_2) = E(\hat{\mu}_3) = \mu$, 所以它们都是总体均数 μ 的无偏估计量。

对于符合无偏性的多个估计量, 究竟哪个估计量更好一些呢? 这可以通过比较它们的方差大小来判别其有效性。

若 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 都是 θ 的无偏估计量, 对任意的样本含量 n , 有 $D(\hat{\theta}_1) \leq D(\hat{\theta}_2)$, 则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$

有效。如果在 θ 的一切无偏估计量中, $\hat{\theta}_1$ 的方差最小, 则称 $\hat{\theta}_1$ 为 θ 的有效估计量。

在上面的例子中, 计算总体均数 μ 的各个无偏估计量的方差可得

$$D(\hat{\mu}_1) = \frac{4}{9}D(X_1) + \frac{1}{9}D(X_2) = \frac{5}{9}D(X) = \frac{5}{9}$$

$$D(\hat{\mu}_2) = \frac{9}{16}D(X_1) + \frac{1}{16}D(X_2) = \frac{5}{8}D(X) = \frac{5}{8}$$

$$D(\hat{\mu}_3) = \frac{1}{4}D(X_1) + \frac{1}{4}D(X_2) = \frac{1}{2}D(X) = \frac{1}{2}$$

由于 $D(\hat{\mu}_3) < D(\hat{\mu}_1) < D(\hat{\mu}_2)$, 所以在三个总体均数 μ 的无偏估计量中, 相对而言, $\hat{\mu}_3$ 为有效估计量。

在有些条件下, 无偏估计量 $\hat{\theta}$ 的方差 $D(\hat{\theta})$ 永远不会小于一个正数 δ , δ 的值可用下式表示

$$\delta = \frac{1}{nE\left[\left(\frac{\partial \ln f(X; \theta)}{\partial \theta}\right)^2\right]}$$

这个正数的值依赖于总体的概率密度, 也依赖于样本的容量 n 。

表示 $D(\hat{\theta})$ 和 δ 的关系式如下

$$D(\hat{\theta}) \geq \frac{1}{nE\left[\left(\frac{\partial \ln f(X; \theta)}{\partial \theta}\right)^2\right]}$$

或

$$D(\hat{\theta}) \geq \frac{1}{nE\left[\left(\frac{\partial^2 \ln f(X; \theta)}{\partial \theta^2}\right)\right]}$$

称上述不等式为罗-克莱姆 (Rao-Gramer) 不等式。

在总体未知参数 θ 的有效估计量存在的条件下, 实际就等同于寻找 $D(\hat{\theta})$ 的下界。当无偏估计量 $\hat{\theta}$ 的方差 $D(\hat{\theta})$ 恰好等于它的下界时, 称它是达到方差界的无偏估计量。

一般而言, 使用任何一个估计量, 都要求样本容量 n 取确定值。但在讨论估计量的有效性时, 应该要考虑 n 无限增大时的大样本性质。能够同时兼顾估计量的偏差性与离散性而建立的最优准则就是要使用极限理论来评估未知参数估计量的一致性。

关于参数估计量的一致性有以下的定义:

设 $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$ 是总体 X 的未知参数 θ 的一系列估计量, 若对于任意的 $\varepsilon > 0$, 恒有 $\hat{\theta}_n$ 以概率收敛于真参数 θ , 即有

$$\lim P\left\{\left|\hat{\theta}_n - \theta\right| \leq \varepsilon\right\} = 1$$

则称 $\hat{\theta}_n$ 为 θ 的一致估计量或相合估计量。

一致估计量要求估计量应随着样本容量 n 的增大而越来越精确。也就是观察值越多, 估计值 $\hat{\theta}$ 应按某种概率意义收敛于 θ 。这说明, 当样本容量充分大时, 估计值越接近于未知参数的真值。

例 4.5 设 X_1, X_2, \dots, X_n 是总体 X 的容量为 n 的样本, $X \sim N(\mu, \sigma^2)$, μ 和 σ^2 是总体 X 的均值和方差, 试证:

(1) \bar{X} 是 μ 的有效估计量;

(2) $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 是总体方差 σ^2 的无偏估计量, 而 $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 是 σ^2 的有偏估计量;

(3) S^2 不是 σ^2 的有效估计量, 但 S^2 是 σ^2 的一致估计量。

证:

(1) 因为

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \cdot n\mu = \mu \end{aligned}$$

所以, \bar{X} 是 μ 的无偏估计量。

又因为, $X \sim N(\mu, \sigma^2)$, 故 X 的概率密度为

$$f(x; \mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < +\infty)$$

取自然对数

$$\ln f(x; \mu) = \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x-\mu)^2}{2\sigma^2}$$

对上式求关于 μ 的一阶和二阶偏导数得

$$\frac{\partial \ln f(x; \mu)}{\partial \mu} = \frac{x - \mu}{\sigma^2}$$

$$\frac{\partial^2 \ln f(x; \mu)}{\partial \mu^2} = -\frac{1}{\sigma^2}$$

对关于 μ 的二阶偏导数求数学期望, 有

$$E\left(\frac{\partial^2 \ln f(x; \mu)}{\partial \mu^2}\right) = E\left(-\frac{1}{\sigma^2}\right) = -\frac{1}{\sigma^2}$$

因此

$$\delta = D(\mu) = \frac{1}{-n\left(-\frac{1}{\sigma^2}\right)} = \frac{\sigma^2}{n}$$

又因为 $D(\bar{X}) = \frac{\sigma^2}{n}$, 所以, $D(\hat{\mu}) = D(\bar{X}) = \sigma^2/n$

因此, \bar{X} 是 μ 的达到方差下界的无偏估计量, 即它是 μ 的有效估计量。

(2) 由于

$$\begin{aligned} E(S^2) &= E\left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right) = \frac{1}{n-1} E\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2)\right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n (D(X_i) + (E(X_i))^2) - n(D(\bar{X}) + (E(\bar{X}))^2)\right) \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] \\ &= \frac{1}{n-1} (n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2) \\ &= \sigma^2 \\ E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right] &= E\left(\frac{n-1}{n} S^2\right) = \frac{n-1}{n} E(S^2) = \frac{n-1}{n} \sigma^2 \end{aligned}$$

所以, 样本方差 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 是总体方差 σ^2 的无偏估计量, 而 $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 是 σ^2 的有偏估计量。当样本含量很大时, 由于 $\frac{n-1}{n} \rightarrow 1$, 所以实际中, 用 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 或 $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 作为总体方差 σ^2 的估计量时, 两者的误差是很小的。

(3) 因为 $X \sim N(\mu, \sigma^2)$, 故 X 的概率密度为

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < +\infty)$$

取自然对数

$$\ln f(x; \mu, \sigma^2) = \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x-\mu)^2}{2\sigma^2} = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

求上式关于 σ^2 的一阶、二阶偏导数得

$$\begin{aligned} \frac{\partial \ln f(x; \mu, \sigma^2)}{\partial \sigma^2} &= \frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \\ \frac{\partial^2 \ln f(x; \mu, \sigma^2)}{\partial (\sigma^2)^2} &= \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6} \end{aligned}$$

所以, 二阶偏导数的数学期望为

$$\begin{aligned} E \left[\frac{\partial^2 \ln f(x; \mu, \sigma^2)}{\partial (\sigma^2)^2} \right] &= \frac{1}{2\sigma^4} - \frac{E[(X-\mu)^2]}{\sigma^6} \\ &= \frac{1}{2\sigma^4} - \frac{1}{\sigma^4} E \left(\frac{X-\mu}{\sigma} \right)^2 \\ &= \frac{1}{2\sigma^4} - \frac{1}{\sigma^4} \left[D \left(\frac{X-\mu}{\sigma} \right) + \left[E \left(\frac{X-\mu}{\sigma} \right) \right]^2 \right] \\ &= \frac{1}{2\sigma^4} - \frac{1}{\sigma^4} [1 + 0^2] \\ &= -\frac{1}{2\sigma^4} \end{aligned}$$

于是

$$\delta = D(\sigma^2) = \frac{1}{-n\left(-\frac{1}{2\sigma^4}\right)} = \frac{2\sigma^4}{n}$$

在第3章中, 我们已经知道 χ^2 的方差 $D(\chi^2) = 2n$, 由此可以求出

$$D(S^2) = \frac{2}{n-1} \sigma^4$$

由于

$$D(S^2) \neq \sigma$$

所以样本方差 S^2 不是总体方差 σ^2 的有效估计量。

根据切比雪夫不等式, 得到

$$P\{|S^2 - \sigma^2| < \varepsilon\} \geq 1 - \frac{D(S^2)}{\varepsilon^2} = 1 - \frac{2\sigma^4}{(n-1)\varepsilon^2}$$

于是

$$\lim_{n \rightarrow \infty} P\{|S^2 - \sigma^2| < \varepsilon\} \geq 1$$

但概率不可能大于1, 所以有

$$\lim_{n \rightarrow \infty} P\{|S^2 - \sigma^2| < \varepsilon\} = 1$$

因此, S^2 是 σ^2 的一致估计量。

判定一个估计量是否为有效估计量, 首先它要满足估计量是无偏估计量, 其次它还要满足达到方差的下界。

在实际应用中, 一个未知参数的估计量并不每次都需要验证这三条性质, 只要满足其中的一两种性质即可。

4.2 参数的区间估计

4.2.1 区间估计的概念

点估计是通过构造样本函数来对总体未知参数进行估计的, 在这种情况下, 用一个样本的观察值计算得到参数的估计值与真值之间的误差到底有多大, 我们是无法知道的。因此, 我们需要在参数空间中给出一个范围, 使待估参数有一个较大概率包含于其内, 这种形式的估计称为参数的区间估计。关于区间估计的确切定义如下:

设总体 X 的分布含有一个未知参数 θ , 若有样本 X_1, X_2, \dots, X_n 确定的两个统计量 $\theta_1(X_1, X_2, \dots, X_n)$ 及 $\theta_2(X_1, X_2, \dots, X_n)$, 对于给定的值 $\alpha (0 < \alpha < 1)$, 满足

$$P\{\theta_1(X_1, X_2, \dots, X_n)\} \leq \theta \leq P\{\theta_2(X_1, X_2, \dots, X_n)\} = 1 - \alpha$$

称随机区间 $[\theta_1, \theta_2]$ 是 θ 的 $100(1-\alpha)\%$ 的置信区间, 称 $1-\alpha$ 为置信度, α 为置信水平或显著性水平, 而把 θ_1 、 θ_2 分别称为置信区间的置信下限和置信上限。

对于由上面定义给出的置信区间, 一般有很多个, 置信区间的大小同置信度 $1-\alpha$ 的大小有关, 一个理想的置信区间, 应是在给定置信度 $1-\alpha$ 下, 区间的平均长度最小的区间。

4.2.2 正态总体均值的置信区间

1. 总体方差 σ^2 已知的情况下

设总体 $X \sim N(\mu, \sigma^2)$, 而 X_1, X_2, \dots, X_n 是从总体 X 中取出的容量为 n 的样本, 在总体方差 σ^2 已知的情况下, 求 X 的期望 μ 的区间估计。

$$\text{由于 } X \sim N(\mu, \sigma^2), \text{ 因此, } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \text{ 而 } U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

由上一节中点估计的理论及其讨论可知, \bar{X} 是 μ 的最好估计量。所以 μ 的区间估计一定与 \bar{X} 有关, \bar{X} 与 μ 之间存在误差是显而易见的, 但只要两者之间的差值的绝对值小于等于一个可以接受的误差限 δ , 即 $|\bar{X} - \mu| \leq \delta$, 则我们就可以得到一个期望 μ 的区间估计 $\bar{X} - \delta \leq \mu \leq \bar{X} + \delta$ 。所以区间估计的关键是如何确定这个可以接受的误差限 δ 。

我们可以把 $|\bar{X} - \mu| \leq \delta$ 作为一个事件来看待, 则当给定一个概率值 $1-\alpha$ 时, δ 的值是可以确定的。

这是因为, 当 $\{P[|\bar{X} - \mu| \leq \delta] = 1 - \alpha\}$ 时, 即

$$P\left\{\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq \frac{\delta}{\sigma/\sqrt{n}}\right\} = 1 - \alpha$$

又

$$P\left\{\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq u_{\alpha/2}\right\} = 1 - \alpha$$

所以, $\frac{\delta}{\sigma/\sqrt{n}} = u_{\alpha/2}$, 得到 $\delta = u_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ 。因此, X 的期望 μ 的区间估计为

$$\left[\bar{X} - u_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right]。$$

置信水平 α 一般取 0.05 或 0.01, 它通常代表 μ 不在估计区间中时的犯错误概率的大小。在 SPSS 中, $u_{\alpha/2}$ 的值可以通过将给定的 $\alpha/2$ 值代入标

准正态分布函数 `IDF.NORMAL(prob, mean, stddev)` 来求得, 其中设置 `mean=0, stddev=1`。称 α 取 0.05 时得到的区间为 95% 的置信区间。

例 4.6 某车间生产滚珠, 从长期实践经验可知滚珠的直径近似服从正态分布, 已知总体的方差为 0.06, 某天从产品中随机抽取 28 个滚珠, 测得直径 (单位: mm) 为

14.6、14.7、15.1、15.0、14.9、15.0、14.8、14.9、14.7、14.8、15.1、15.0、15.2、15.1、14.8、14.7、15.0、14.9、14.8、15.1、15.0、14.9、15.1、15.2、14.9、14.8、14.9、14.8
求平均直径 $\alpha = 0.05$ 的置信区间。

在 SPSS 中的解题步骤:

- ① 将题中数据在 SPSS 的数据编辑窗口中制成数据文件, 见 data04-03.sav。
- ② 按 **Analyze**→**Reports**→**Case Summaries** 顺序, 打开 Case Summaries 对话框 (见图 2-51)。

在左边的变量名源框中, 选中滚珠直径变量, 通过中间的右移箭头将它移到 **Variables** 框中。

- ③ 单击 **Statistics** 按钮, 打开 **Statistics** 选项卡 (见图 2-52)。

在 **Statistics** 对话框中选择 **Mean** 统计量, 并将其移入到 **Cell Statistics** 下框中。

单击 **Continue** 按钮, 返回 **Case Summaries** 对话框。

- ④ 单击 **OK** 按钮, 在输出窗口中得到样本的平均数的计算结果, 见表 4-4。

⑤ 按 **File**→**New**→**Data** 顺序, 打开数据编辑窗口, 新建一个数据文件 data04-04.sav。它有 4 个变量, 分别为滚珠直径均值 (设置其小数点后保留 4 位)、总体方差、显著性水平、样本含量, 并依次将 14.9214、0.06、0.05 和 28 输入到这些变量下。

⑥ 按 **Transform**→**Compute Variable** 顺序, 打开 **Compute Variable** 对话框 (见图 2-53)。在 **Target Variable** 框中, 输入目标变量名为 **下限**。在 **Numeric Expression** 框中, 输入 “滚珠直径均值-IDF.NORMAL(1-显著性水平/2, 0,1)*sqrt(总体方差/样本含量)”。

⑦ 单击 **OK** 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现下限的新变量及其值。

⑧ 按 **Transform**→**Compute Variable** 顺序, 打开 **Compute Variable** 对话框 (见图 2-53)。在 **Target Variable** 框中, 输入目标变量名为 **上限**。在 **Numeric Expression** 框中, 输入 “滚珠直径均值+IDF.NORMAL(1-显著性水平/2, 0,1)*sqrt(总体方差/样本含量)”。

⑨ 单击 **OK** 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现上限的新变量及其值。见图 4-1。

表 4-4 均值

Case Summaries

Mean

滚珠直径
14.9214

滚珠直径均值	总体方差	显著性水平	样本含量	下限	上限
14.9124	0.06	0.05	28.00	14.82	15.00

图 4-1 平均直径 $\alpha = 0.05$ 的置信区间

因此，滚珠平均直径 95% ($\alpha = 0.05$) 的置信区间为 [14.82, 15.00]。

2. 总体方差 σ^2 未知的情况下

在总体方差 σ^2 未知的情况下，可以用样本的方差 S^2 来作为总体方差 σ^2 的无偏估计量。由于

$$\begin{aligned}
 S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n [(X_i - \mu) - (\bar{X} - \mu)]^2 \\
 &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i - \mu)^2 - \frac{\left[\sum_{i=1}^n (\bar{X} - \mu) \right]^2}{n} \right] \\
 &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \right]
 \end{aligned}$$

因此可得

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} - \frac{n(\bar{X} - \mu)^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 - \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2$$

因为 $\frac{X_i - \mu}{\sigma} \sim N(0,1)$ ， X_1, X_2, \dots, X_n 相互独立，从而 $\frac{X_1 - \mu}{\sigma}, \frac{X_2 - \mu}{\sigma}, \dots, \frac{X_{n-1} - \mu}{\sigma}$ 相互独立，根据第 3 章中 χ^2 分布定义可知， $\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$ 。又， $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ ，所以， $\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2 \sim \chi^2(1)$ ，故， $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ 。

又因为 \bar{X} 与 S^2 独立，所以根据第 3 章中 t 分布定义可得

$$T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \bigg/ \sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

因此，参照总体方差已知时的做法可得

$$P\left\{\frac{|\bar{X}-\mu|}{S/\sqrt{n}} \leq t_{\alpha/2}(n-1)\right\} = 1-\alpha$$

即

$$P\left\{\bar{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right\} = 1-\alpha$$

所以, 总体均值 μ 的 $100(1-\alpha)\%$ 的置信区间为

$$\left[\bar{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \bar{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right],$$

其中的 $t_{\alpha/2}(n-1)$ 的值可以通过 SPSS 中的 $IDF.T(\text{prob}, \text{df})$ 函数来获得。

例 4.7 随机地从一批零件中抽取 34 个, 分别测得其长度 (单位: cm), 并将结果存放在 data04-05.sav 中, 从经验知该零件长度近似服从正态分布, 试求其均值 90% 的置信区间。

在 SPSS 中求本例所要求的置信区间的方法为:

① 在 SPSS 数据编辑窗口中, 打开数据文件 data04-05.sav。

② 按 Analyze→Descriptive Statistics→Explore 顺序, 打开 Explore 对话框, 见图 2-99。在左边的变量名源框中, 选中零件长度变量, 通过中间的右移箭头将它们移到 Dependent List 框中, 单击 Statistics 按钮, 打开 Statistics 对话框, 见图 2-100。

③ 在 Statistics 对话框中, 选择 Descriptives 选项, 在 Confidence Interval for Mean: 后框中输入 90。单击 Continue 按钮返回 Explore 对话框。其他保持系统默认选择。

④ 单击 OK 按钮运行, 在输出窗口中得到计算结果, 见表 4-5。

⑤ 结果说明

所求零件长度的 90% 的置信区间为: [2.1202, 2.1298]。

表 4-5 90% 置信区间的计算结果

Descriptives				Statistic	Std. Error
零件长度	Mean			2.1250	.00281
	90% Confidence Interval for Mean	Lower Bound		2.1202	
		Upper Bound		2.1298	
	5% Trimmed Mean			2.1250	
	Median			2.1300	
	Variance			.000	
	Std. Deviation			.01638	
	Minimum			2.10	
	Maximum			2.15	
	Range			.05	
	Interquartile Range			.03	
	Skewness			-.176	.403
	Kurtosis			-1.059	.788

4.2.3 正态总体方差的置信区间

因为, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, 所以, 令 $P\left\{a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right\} = 1-\alpha$ 。考虑到 χ^2 是偏态

分布, 为简化计算, 可令, $P\left\{\frac{(n-1)S^2}{\sigma^2} < a\right\} = P\left\{\frac{(n-1)S^2}{\sigma^2} > b\right\} = \frac{\alpha}{2}$, 故得 $a = \chi_{\alpha/2}^2(n-1)$, $b = \chi_{1-\alpha/2}^2(n-1)$ 。

注: α 为 χ^2 左侧累积概率。

由 $\chi_{\alpha/2}^2(n-1) \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{1-\alpha/2}^2(n-1)$, 可解得 σ^2 的 $100(1-\alpha)\%$ 的置信区间为

$$\left[\frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)}, \frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)} \right]$$

其中的 $\chi_{\alpha/2}^2(n-1)$ 和 $\chi_{1-\alpha/2}^2(n-1)$ 的值可以通过 SPSS 中的 IDF.CHISQ (prob, df) 函数来获得。

例 4.8 已知某种木材横纹抗压力的实验值服从正态分布 $N(\mu, \sigma^2)$, 对 20 根木材作横纹抗压力试验的结果存放在 data04-06.sav 中, 使对该木材横纹抗压力的方差进行区间估计 ($\alpha = 0.05$)。

在 SPSS 中本例的解题步骤为:

① 在 SPSS 的数据编辑窗口中, 打开数据文件 data04-06.sav。

② 按 Analyze→Descriptive Statistics→Descriptives 顺序, 展开 Descriptives 对话框, 见图 2-95。

在左侧变量名源框中选择抗压力值, 将其移入到 Variable(s) 下框中。

③ 单击 Options 按钮, 弹出 Options 对话框, 见图 2-96。在 Options 对话框中, 选择 Variance 选项, 按 Continue 返回 Descriptives 对话框。

表 4-6 样本方差计算结果

Descriptive Statistics

	N	Variance
抗压力值	20	1.003E3
Valid N(Listwise)	20	

④ 单击 OK 按钮执行, 在输出窗口中, 出现计算结果, 见表 4-6。

从表 4-6 中可见, 样本方差为 1003。

⑤ 按 File→New→Data 顺序, 打开数据编辑窗口, 新建一个数据文件 data04-07.sav。它有 3 个变量, 分别为方差、显著性水平和样本含量,

并依次将 1003、0.05 和 20 输入到这些变量下。

⑥ 按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框(见图 2-53)。在 Target Variable 框中, 输入目标变量名为 下限。在 Numeric Expression 框中, 输入“(样本含量-1)*方差/IDF.CHISQ(1-显著性水平/2, 样本含量-1)”。

⑦ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 下限的新变量

及其值。

⑧ 按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框(见图 2-53)。在 Target Variable 框中, 输入目标变量名为 *上限*。在 Numeric Expression 框中, 输入“(样本含量-1)*方差/IDF.CHISQ(显著性水平/2,样本含量-1)”。

⑨ 按 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 *上限* 的新变量及其值。见图 4-2。

方差	显著性水平	样本含量	下限	上限
1003.00	0.05	20	580.08	2139.67

图 4-2 方差 95% 的置信区间

因此, 该木材横纹抗压力的方差 95% 的置信区间为 $[580.08, 2139.67]$ 。

4.2.4 两个正态总体均值差和方差比的区间估计

为做出二正态总体均值差和方差比的区间估计, 首先设两个总体 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 而 X_1, X_2, \dots, X_{n_1} , Y_1, Y_2, \dots, Y_{n_2} 分别是来自总体 X 和 Y 中取出的容量为 n_1 和 n_2 的样本, 且总体 X 和 Y 的样本均值分别为

$$\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i, \quad \bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$$

它们分别是总体均值 μ_1 和 μ_2 的无偏估计值, 总体 X 和 Y 的样本方差分别为

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

它们分别是总体方差 σ_1^2 和 σ_2^2 的无偏估计值。

4.2.4.1 两个正态总体均值差的区间估计

1. 总体方差 σ_1^2 和 σ_2^2 已知时的 $\mu_1 - \mu_2$ 的区间估计

由于两个样本 X_1, X_2, \dots, X_{n_1} 与 Y_1, Y_2, \dots, Y_{n_2} 分别来自两个独立的总体 X 与 Y, 所以两个样本之间互相独立, 故 \bar{X} 与 \bar{Y} 相互独立。又由于

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2, \quad D(\bar{X} - \bar{Y}) = D(\bar{X}) + D(\bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

因而

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

因此可得

$$U = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

对于给定的置信度 $1-\alpha$ ，由 $P\{-u_{\alpha/2} \leq U \leq u_{\alpha/2}\} = 1-\alpha$ ，作等价变换得

$$P\left\{(\bar{X} - \bar{Y}) - u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{X} - \bar{Y}) + u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\} = 1-\alpha$$

由此可得 $\mu_1 - \mu_2$ 的 $100(1-\alpha)\%$ 的置信区间为

$$\left[(\bar{X} - \bar{Y}) - u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{X} - \bar{Y}) + u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

例 4.9 1986 年在某地区对职工平均工资情况进行了分行业调查，已知体育、卫生、社会福利事业职工 X （单位：元） $\sim N(\mu_1, 218^2)$ ；文教、艺术、广播事业职工工资 Y （单位：元） $\sim N(\mu_2, 227^2)$ ，从总体 X 中随机调查了 25 人，得 $\bar{X} = 1286$ 元，从总体 Y 中随机调查了 30 人，得 $\bar{Y} = 1272$ 元，求这两大行业职工平均工资之差的 95% 的置信区间。

① 在 SPSS 的数据编辑窗口中，建立数据文件 data04-08.sav。其中所建变量及输入的数值见图 4-3。

第一类样本均值	第二类样本均值	样本1容量	第一类总体标准差	第二类总体标准差	样本2容量	显著性水平
1286.00	1272.00	25.00	218.00	227.00	30.00	0.05

图 4-3 data04-08.sav 中的变量及数据值

② 按 Transform→Compute Variable 顺序，打开 Compute Variable 对话框（见图 2-53）。在 Target Variable 框中，输入目标变量名为 下限。在 Numeric Expression 框中，输入“第一类样本均值-第二类样本均值-IDF.NORMAL(1-显著性水平/2,0,1)*sqrt(第一类总体标准差*第一类总体标准差/样本 1 容量+第二类总体标准差*第二类总体标准差/样本 2 容量)”。

③ 单击 OK 按钮运行，则在数据编辑窗口工作的数据文件中，出现 下限 的新变量及其值。

④ 按 Transform→Compute Variable 顺序，打开 Compute Variable 对话框（见图 2-53）。在 Target Variable 框中，输入目标变量名为 上限。在 Numeric Expression 框中，输入“第一类样本均值-第二类样本均值-IDF.NORMAL(1-显著性水平/2,0,1)*sqrt(第一类总体标准差*第一类总体标准差/样本 1 容量+第二类总体标准差*第二类总体标准差/样本 2 容量)”。

⑤ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现上列的新变量及其值。见图 4-4。

第一类样本均值	第二类样本均值	样本1容量	第一类总体标准差	第二类总体标准差	样本2容量	显著性水平	下限	上限
1286.00	1272.00	25.00	218.00	227.00	30.00	0.05	-103.90	131.90

图 4-4 两总体均值差 95% 的置信区间

因此, 两总体均值差 95% 的置信区间为 $[-103.90, 131.90]$ 。

2. 总体方差 σ_1^2 和 σ_2^2 未知但 $\sigma_1^2 = \sigma_2^2$ 时的 $\mu_1 - \mu_2$ 的区间估计

参照 4.1.3 的 4 (1) ①中的推理过程可得

$$U = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

而

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

因此根据 t 分布的定义可得

$$\begin{aligned} T &= \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \bigg/ \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2}} \\ &= \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2) \end{aligned}$$

则在给定的置信度 $1 - \alpha$ 下, 仿 4.1.3 的 4 (1) ①中方法可得 $\mu_1 - \mu_2$ 的 $100(1 - \alpha)\%$ 的置信区间为

$$\begin{aligned} &\left[(\bar{X} - \bar{Y}) - t_{\alpha/2}(n_1 + n_2 - 2) \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \right. \\ &\left. (\bar{X} - \bar{Y}) + t_{\alpha/2}(n_1 + n_2 - 2) \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right] \end{aligned}$$

例 4.10 某种子分公司经过试验获得 A、B 两种杂交玉米产量的数据如下:

品种 A: 86、87、56、93、84、93、75、79、80

品种 B: 80、79、58、91、77、82、74、66

经检验两种杂交玉米产量均服从正态分布, 且方差相等, 求总体均值差 $\mu_A - \mu_B$ 的 95% 的置信区间。

① 在 SPSS 的数据编辑窗口中, 建立数据文件, 见 data04-09.sav。

② 按 Analyze → Compare Means → Independent-Samples T Test 顺序, 展开 Independent-Samples T Test 对话框, 见图 4-5。

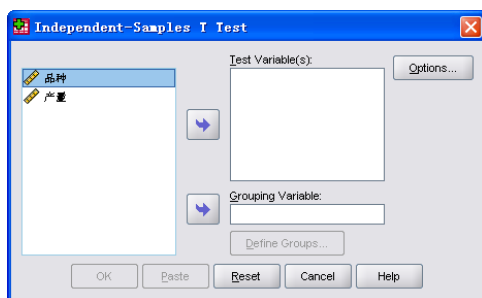


图 4-5 Independent-Samples T Test 对话框

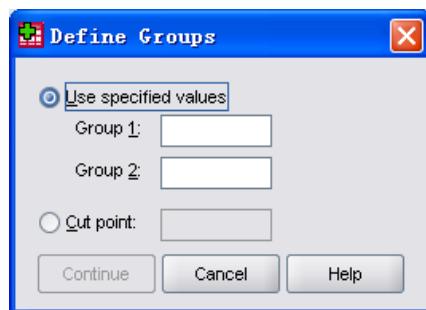


图 4-6 Define Groups 对话框

在变量名框中将产量移入到 Test Variable(s)下框中, 将品种移入到 Grouping Variable 下框中。

③ 单击 Define Groups, 弹出 Define Groups 对话框, 见图 4-6。

在 Group1 中输入 1, 在 Group2 中输入 2, 或选择 Cut point, 并在其后框中输入 1.5。单击 Continue 返回 Independent-Samples T Test 对话框。

④ 单击 OK 按钮, 则在输出窗中得到想要的结果, 见表 4-7。

表 4-7 两正态总体方差未知且相等时均值差的 95% 的置信区间

Independent Samples Test									
	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
产量	.026	.875	1.065	15	.304	5.56944	5.22907	-5.57606	16.71495
Equal variances Assumed			1.073	14.997	.300	5.56944	5.19201	-5.49651	16.66740
Equal variances not Assumed									

由表 4-7 所标的框中的值即是我们所要求的两总体均值差的 95% 的置信区间, 也就是 $[-5.57606, 16.71495]$ 。

关于本表中涉及的其他内容, 将在第 6 章正态总体均值差异的显著性检验中详述。

如果本例中要求 95% 以外的其他置信区间, 则只需要 Independent-Samples T Test 对话框中单击 Options 按钮, 在弹出的 Options 对话框 (见图 4-7) 中, 在 Confidence Interval 后框中作相应修改即可, 系统默认值为 95, 即求 95% 的置信区间。

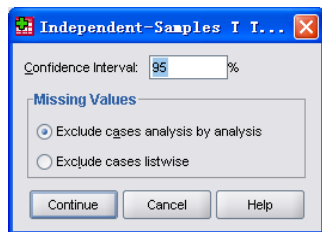


图 4-7 Options 对话框

本例也可用前面已介绍过的方法, 在先求两样本的均值和方差后, 再按下例所示进行计算, 但在已知原始数据的前提下, 无疑现在的做法是最简捷、省时的。

例 4.11 设小麦品种 1 的生长天数 $X \sim N(\mu_1, \sigma^2)$, 小麦品种 2 的生长天数 $Y \sim N(\mu_2, \sigma^2)$ 。现从 X 中随机抽取容量为 10 的样本, 得 $\bar{x} = 99.2$, $s_1^2 = 0.84$; 从 Y 中随机抽取容量为 10 的样本, 得 $\bar{y} = 98.9$, $s_2^2 = 0.77$, 求 $\mu_1 - \mu_2$ 的 95% 的置信区间。

① 在 SPSS 的数据编辑窗口中, 建立数据文件 data04-10.sav。其中所建变量及输入的数值见图 4-8。

样本1均值	样本1方差	样本1容量	样本2均值	样本2方差	样本2容量
99.20	0.84	10.00	98.90	0.77	10.00

图 4-8 data04-10.sav 中的变量及数据值

② 按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 下限。在 Numeric Expression 框中, 输入 “样本 1 均值-样本 2 均值-IDF.T(1-0.05/2, 样本 1 容量+样本 2 容量-2)*sqrt(((样本 1 容量-1)*样本 1 方差+(样本 2 容量-1)*样本 2 方差))*sqrt((样本 1 容量+样本 2 容量)/(样本 1 容量*样本 2 容量*(样本 1 容量+样本 2 容量-2)))”。

③ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 下限 的新变量及其值。

④ 按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 上限。在 Numeric Expression 框中, 输入 “样本 1 均值-样本 2 均值+IDF.T(1-0.05/2, 样本 1 容量+样本 2 容量-2)*sqrt(((样本 1 容量-1)*样本 1 方差+(样本 2 容量-1)*样本 2 方差))*sqrt((样本 1 容量+样本 2 容量)/(样本 1 容量*样本 2 容量*(样本 1 容量+样本 2 容量-2)))”。

⑤ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 上限 的新变量及

其值。见图 4-9。

样本1均值	样本1方差	样本1容量	样本2均值	样本2方差	样本2容量	下限	上限
99.20	0.84	10.00	98.90	0.77	10.00	-0.54	1.14

图 4-9 两正态总体方差未知相等时均值差 95% 的置信区间

因此，两正态总体方差未知但相等时均值差 95% 的置信区间为 $[-0.54, 1.14]$ 。

3. 总体方差 σ_1^2 和 σ_2^2 未知但 $\sigma_1^2 \neq \sigma_2^2$ 时的 $\mu_1 - \mu_2$ 的区间估计

在大样本情况下，即 n_1, n_2 均较大时，由中心极限定理可得

$$U = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \xrightarrow{L} N(0, 1)$$

可以证明，当用 S_1^2, S_2^2 分别替代 σ_1^2, σ_2^2 后，仍有相同的结论。因此，可得 $\mu_1 - \mu_2$ 的近似的 $100(1-\alpha)\%$ 的置信区间为

$$\left[(\bar{X} - \bar{Y}) - u_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{X} - \bar{Y}) + u_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]$$

在 n_1, n_2 均较小时，统计量

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(f)$$

其中自由度

$$f = \left\{ \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2 \middle/ \left[\frac{(S_1^2/n_1)^2}{n_1 + 1} + \frac{(S_2^2/n_2)^2}{n_2 + 1} \right] \right\} - 2$$

当 f 不为正整数时，对 f 取整加 1，作为最终的自由度。

此时，可得 $\mu_1 - \mu_2$ 的近似的 $100(1-\alpha)\%$ 的置信区间为

$$\left[(\bar{X} - \bar{Y}) - t_{\alpha/2}(f) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{X} - \bar{Y}) + t_{\alpha/2}(f) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]$$

例 4.12 用两种除草剂作除草效果试验，各喷洒 5 个试验小区，除草的杂草数如下：
除草剂 A：9、6、12、7、7

除草剂 B: 177、151、110、117、135

经检验两种除草剂的除草数均服从正态分布，但方差不相等，求总体均值差 $\mu_A - \mu_B$ 的 95% 的置信区间。

① 在 SPSS 的数据编辑窗口中，建立数据文件，见 data04-11.sav。

② 按 Analyze→Compare Means→Independent-Samples T Test 顺序，展开 Independent-Samples T Test 对话框，见图 4-5。

在变量名框中将除草数移入到 Test Variable(s) 下框中，将除草剂类型移入到 Grouping Variable 下框中。

③ 单击 Define Groups，弹出 Define Groups 选项卡，见图 4-6。

在 Group1 中输入 1，在 Group2 中输入 2，或选择 Cut point，并在其后框中输入 1.5。按 Continue 返回 Independent-Samples T Test 对话框。

④ 单击 OK 按钮，则在输出窗中得到想要的结果，见表 4-8。

表 4-8 两正态总体方差未知但不相等时均值差的 95% 的置信区间

Independent Samples Test									
	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
产量	9.378	.016	-10.693	8	.000	-129.800000	12.13837	-157.79113	-101.80847
Equal variances			-10.693	4.062	.000	-129.80000	12.13837	-163.29841	-36.30159
Assumed									
Equal variances not Assumed									

由表 4-7 所标的框中的值即是我们所要求的两总体均值差的 95% 的置信区间，也就是 $[-163.29841, -96.30159]$ 。

例 4.13 在验证尿蛋白的排出量的多少可以用来反映运动量大小的试验中，研究人员选取两组原始条件相似的运动员作如下试验：一组 18 人做一般运动量练习，结果测得其运动后的尿蛋白排出量为： $\bar{x}_1 = 14.7 \mu\text{g/ml}$, $S_1^2 = 6.25(\mu\text{g/ml})^2$ ，另一组 17 人做大量运动量练习，结果测得其运动后的尿蛋白排出量为： $\bar{x}_2 = 39.3 \mu\text{g/ml}$, $S_2^2 = 26.01(\mu\text{g/ml})^2$ ，试求两总体尿蛋白排出量均值之差的 95% 的置信区间。

① 在 SPSS 的数据编辑窗口中，建立数据文件 data04-12.sav。其中所建变量及输入的数值见图 4-10。

对照组均值	对照组方差	对照组容量	实验组均值	实验组方差	实验组容量	显著性水平
14.70	6.25	18.00	39.30	26.01	17.00	0.05

图 4-10 data04-12.sav 中的变量及数据值

② 计算自由度 f

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 f 。在 Numeric Expression 框中, 输入“TRUNC((对照组方差/对照组容量+实验组方差/实验组容量)*(对照组方差/对照组容量+实验组方差/实验组容量)/((对照组方差/对照组容量)*(对照组方差/对照组容量)/(对照组容量+1)+(实验组方差/实验组容量)*(实验组方差/实验组容量)/(实验组容量+1))-2)+1”。

③ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 f 的新变量及其值。

④ 计算下限

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 下限。在 Numeric Expression 框中, 输入“对照组均值-实验组均值-IDF.T(1-显著性水平/2, f)*sqrt(对照组方差/对照组容量+实验组方差/实验组容量)”。

⑤ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现下限的新变量及其值。见图 4-11。

⑥ 计算上限

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 上限。在 Numeric Expression 框中, 输入“对照组均值-实验组均值+IDF.T(1-显著性水平/2, f)*sqrt(对照组方差/对照组容量+实验组方差/实验组容量)”。

⑦ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现上限的新变量及其值。见图 4-11。

对照组均值	对照组方差	对照组容量	实验组均值	实验组方差	实验组容量	显著性水平	f	下限	上限
14.70	6.25	18.00	39.30	26.01	17.00	0.05	24.00	-27.43	-21.77

图 4-11 两正态总体方差未知且不相等时均值差 95% 的置信区间

因此, 两正态总体方差未知且不相等时均值差 95% 的置信区间为 $[-27.43, -21.77]$ 。

4.2.4.2 两个正态总体方差比的区间估计

由于 $\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1)$, $\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$, 同时两个样本相互独立, 意味着 S_1^2

与 S_2^2 相互独立, 根据 F 分布的定义可得

$$F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2} / (n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2} / (n_2-1)} = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

设 $\frac{S_1^2}{S_2^2} \geq 1$, 则在给定的置信度 $1-\alpha$ 下, 有

$$P\{F_{\alpha/2}(n_1-1, n_2-1) \leq F \leq F_{1-\alpha/2}(n_1-1, n_2-1)\} = 1-\alpha$$

即

$$P \leq \left\{ \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{1-\alpha/2}(n_1-1, n_2-1)} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{\alpha/2}(n_1-1, n_2-1)} \right\} = 1-\alpha$$

因此 $\frac{\sigma_1^2}{\sigma_2^2}$ 的 $100(1-\alpha)\%$ 的置信区间为

$$\left[\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{1-\alpha/2}(n_1-1, n_2-1)}, \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{\alpha/2}(n_1-1, n_2-1)} \right]$$

例 4.14 仍以例 4.13 中数据为例, 试求两总体尿蛋白排出量方差比的 95% 的置信区间。

① 在 SPSS 的数据编辑窗口中, 打开数据文件 data04-12.sav。

② 计算两总体方差比下限

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为方差比下限。在 Numeric Expression 框中, 输入“对照组方差/(实验组方差*IDF.F(1-显著性水平/2, 对照组容量-1, 实验组容量-1))”。

③ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现方差比下限的新变量及其值。见图 4-12。

④ 计算两总体方差比上限

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为方差比上限。在 Numeric Expression 框中, 输入“对照组方差/(实验组方差*IDF.F(显著性水平/2, 对照组容量-1, 实验组容量-1))”。

⑤ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现方差比上限的新变量及其值。见图 4-12。

对照组均值	对照组方差	对照组容量	实验组均值	实验组方差	实验组容量	显著性水平	f	下限	上限	方差比下限	方差比上限
14.70	6.25	18.00	39.30	26.01	17.00	0.05	24.00	-27.43	-21.77	0.09	0.65

图 4-12 两正态总体方差比 95% 的置信区间

因此, 两正态总体方差比的 95% 的置信区间为 $[0.09, 0.65]$ 。

4.2.5 非正态总体参数的近似区间估计

4.2.5.1 二项分布总体参数的区间估计

设 X_1, X_2, \dots, X_n 是来自总体 $X \sim B(1, p)$ 的一个大样本, 则 $\sum_{i=1}^n X_i \sim B(n, p)$, 而样本均值

$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 的概率分布为

$$P\left\{\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{j}{n}\right\} = C_n^j p^j q^{n-j} \quad (j=0, 1, 2, \dots, n)$$

其中 $q=1-p$

由此可得

$$E(\bar{X}) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} \cdot np = p$$

$$D(\bar{X}) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} D\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} npq = \frac{pq}{n}$$

根据棣莫弗—拉普拉斯中心极限定理可知

$$\bar{X} \sim N\left(p, \frac{p(1-p)}{n}\right) \quad (np > 5, n(1-p) > 5)$$

因此

$$U = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

对于给定的置信度 $1-\alpha$, 由 $P\{-u_{\alpha/2} \leq U \leq u_{\alpha/2}\} = 1-\alpha$, 作等价变换得

$$|\bar{X} - p| \leq u_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

即

$$(n + \mu_{\alpha/2}^2)p^2 - (2n\bar{X} + \mu_{\alpha/2}^2)p + n\bar{X}^2 \leq 0$$

解上述不等式得

$$\begin{cases} p_1 = \frac{(2n\bar{X} + \mu_{\alpha/2}^2) - \sqrt{(2n\bar{X} + \mu_{\alpha/2}^2)^2 - 4(2n\bar{X} + \mu_{\alpha/2}^2)n\bar{X}^2}}{2(n + \mu_{\alpha/2}^2)} \\ p_2 = \frac{(2n\bar{X} + \mu_{\alpha/2}^2) + \sqrt{(2n\bar{X} + \mu_{\alpha/2}^2)^2 - 4(2n\bar{X} + \mu_{\alpha/2}^2)n\bar{X}^2}}{2(n + \mu_{\alpha/2}^2)} \end{cases}$$

令 $a = n + \mu_{\alpha/2}^2$, $b = 2n\bar{X} + \mu_{\alpha/2}^2$, $c = n\bar{X}^2$, 则上式可以简写为

$$\begin{cases} p_1 = \frac{b - \sqrt{b^2 - 4ac}}{2a} \\ p_2 = \frac{b + \sqrt{b^2 - 4ac}}{2a} \end{cases}$$

故 p 的 $100(1-\alpha)\%$ 的置信区间为 $[p_1, p_2]$ 。

例 4.15 为检验某体育用品厂生产的羽毛球的合格率, 从该厂生产的一大批羽毛球中随机抽取 150 只羽毛球进行检验, 达到一级品的有 107 只, 求该批羽毛球达到一级品率的 95% 的置信区间。

① 在 SPSS 的数据编辑窗口中, 建立数据文件 data04-13.sav。其中所建变量及输入的数值见图 4-13。

试验次数	出现次数
150.00	107.00

图 4-13 data04-13.sav 中的变量及数据值

② 计算 a

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框。在 Target Variable 框中, 输入目标变量名为 a 。在 Numeric Expression 框中, 输入 “试验次数+IDF.NORMAL(1-0.05/2,0,1)*IDF.NORMAL(1-0.05/2,0,1)”。

③ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 a 的新变量及其值。

④ 计算 b

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框。在 Target Variable 框中, 输入目标变量名为 b 。在 Numeric Expression 框中, 输入 “2*出现次数+IDF.NORMAL(1-0.05/2,0,1)*IDF.NORMAL(1-0.05/2,0,1)”。

⑤ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 b 的新变量及其值。

⑥ 计算 c

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框。在 Target Variable 框中, 输入目标变量名为 c 。在 Numeric Expression 框中, 输入“试验次数*(出现次数/试验次数)*(出现次数/试验次数)”。

⑦ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 c 的新变量及其值。

⑧ 计算下限 $p1$

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 $p1$ 。在 Numeric Expression 框中, 输入“ $1/(2*a)*(b-\sqrt{b*b-4*a*c})$ ”。

⑨ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 $p1$ 的新变量及其值。见图 4-14。

⑩ 计算上限 $p2$

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 $p2$ 。在 Numeric Expression 框中, 输入“ $1/(2*a)*(b+\sqrt{b*b-4*a*c})$ ”。

⑪ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 $p2$ 的新变量及其值。见图 4-14。

试验次数	出现次数	a	b	c	p1	p2
150.00	107.00	153.84	217.84	76.33	0.64	0.78

图 4-14 两正态总体方差比的 95% 的置信区间

因此, 该批羽毛球达到一级品率的 95% 的置信区间为 [0.64, 0.78]。

4.2.5.1 两个二项分布总体参数的区间估计

设 X_1, X_2, \dots, X_{n_1} 与 Y_1, Y_2, \dots, Y_{n_2} 是分别从总体 $X \sim B(1, p_1), Y \sim B(1, p_2)$ 中随机抽取的两个相互独立的样本, 且 $n_1 p_1 > 5, n_1(1-p_1) > 5, n_2 p_2 > 5, n_2(1-p_2) > 5$, 因为

$$\bar{X} \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \bar{Y} \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

所以

$$\bar{X} - \bar{Y} \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

因此

$$U = \frac{\bar{X} - \bar{Y} - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$$

由此可得 $p_1 - p_2$ 的 $100(1-\alpha)\%$ 的置信区间为

$$\left[(\bar{X} - \bar{Y}) - \mu_{\alpha/2} \sqrt{\frac{\bar{X}_1(1-\bar{X}_1)}{n_1} + \frac{\bar{Y}_2(1-\bar{Y}_2)}{n_2}}, (\bar{X} - \bar{Y}) + \mu_{\alpha/2} \sqrt{\frac{\bar{X}_1(1-\bar{X}_1)}{n_1} + \frac{\bar{Y}_2(1-\bar{Y}_2)}{n_2}} \right]$$

例 4.16 为检查两种小麦的抗虫性,从品种 A 小麦中随机抽取 300 穗受虫害的有 63 穗,从品种 B 小麦中随机抽取 260 穗受虫害的有 42 穗,试求 $p_A - p_B$ 的 95% 的置信区间。

① 在 SPSS 的数据编辑窗口中,建立数据文件 data04-14.sav。其中所建变量及输入的数值见图 4-15。

查A数量	A虫害数量	查B数量	B虫害数量
300.00	63.00	260.00	42.00

图 4-15 data04-14.sav 中的变量及数据值

② 计算 PA

按 Transform→Compute Variable 顺序,打开 Compute Variable 对话框。在 Target Variable 框中,输入目标变量名为 PA 。在 Numeric Expression 框中,输入“A 虫害数量/查 A 数量”。

③ 单击 OK 按钮运行,则在数据编辑窗口工作的数据文件中,出现 PA 的新变量及其值。

④ 计算 PB

按 Transform→Compute Variable 顺序,打开 Compute Variable 对话框。在 Target Variable 框中,输入目标变量名为 PB 。在 Numeric Expression 框中,输入“B 虫害数量/查 B 数量”。

⑤ 单击 OK 按钮运行,则在数据编辑窗口工作的数据文件中,出现 PB 的新变量及其值。

⑥ 计算下限

按 Transform→Compute Variable 顺序,打开 Compute Variable 对话框(见图 2-53)。在 Target Variable 框中,输入目标变量名为 下限。在 Numeric Expression 框中,输入“ $PA - PB - \text{IDF.NORMAL}(1-0.05/2,0,1) * \sqrt{PA * (1-PA) / \text{查 A 数量} + PB * (1-PB) / \text{查 B 数量}}$ ”。

⑦ 单击 OK 按钮运行,则在数据编辑窗口工作的数据文件中,出现 下限 的新变量及

其值。见图 4-16。

⑧ 计算上限

按 Transform→Compute Variable 顺序，打开 Compute Variable 对话框（见图 2-53）。在 Target Variable 框中，输入目标变量名为上限。在 Numeric Expression 框中，输入“PA-PB+IDF.NORMAL(1-0.05/2,0,1)*sqrt(PA*(1-PA)/查 A 数量+PB*(1-PB)/查 B 数量)”。

⑨ 单击 OK 按钮运行，则在数据编辑窗口工作的数据文件中，出现上限的新变量及其值。见图 4-16。

查A数量	A虫害数量	查B数量	B虫害数量	PA	PB	下限	上限
300.00	63.00	260.00	42.00	0.21	0.16	-0.0158	0.1127

图 4-16 两个二项分布总体参数差的 95% 的置信区间

因此，本例所求的 $PA - PB$ 的 95% 的置信区间为 $[-0.0158, 0.1127]$ 。

4.2.6 其他总体参数及参数的区间估计

1. 当总体分布类型未知， $E(X) = \mu$ 未知，而 $D(X) = \sigma^2$ 已知时，求总体均数 μ 的区间估计，可以从该总体 X 中随机抽取一个大样本 X_1, X_2, \dots, X_n ，则由中心极限定理可知

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

因此，按 4.1.3 中 2 (1) 里的做法， X 的期望 μ 的 $100(1-\alpha)\%$ 的置信区间为

$$\left[\bar{X} - u_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

例 4.17 某车间用一台包装机包装精盐，额定标准每袋净重 500g， $\sigma = 15g$ ，某天随机地抽取 80 袋，称得净重的均值为 $\bar{x} = 499.6g$ ，试求该包装机包装精盐的均值 95% 的置信区间。

样本均值	样本容量	总体标准差
499.60	80.00	15.00

图 4-17 data04-15.sav 中的变量及数据值

① 在 SPSS 的数据编辑窗口中，建立数据文件 data04-15.sav。其中所建变量及输入的数值见图 4-17。

② 计算下限

按 Transform→Compute Variable 顺序，打开 Compute Variable 对话框（见图 2-53）。在 Target Variable 框中，输入目标变量名为下限。在 Numeric Expression 框中，输入“样本均值-IDF.NORMAL(1-0.05/2,0,1)*总体标准差/sqrt(样本容量)”。

③ 单击 OK 按钮运行,则在数据编辑窗口工作的数据文件中,出现 下限 的新变量及其值。见图 4-18。

④ 计算上限

按 Transform→Compute Variable 顺序,打开 Compute Variable 对话框(见图 2-53)。在 Target Variable 框中,输入目标变量名为 上限。在 Numeric Expression 框中,输入“样本均值+IDF.NORMAL(1-0.05/2,0,1)*总体标准差/sqrt(样本容量)”。

⑤ 单击 OK 按钮运行,则在数据编辑窗口工作的数据文件中,出现 上限 的新变量及其值。见图 4-18。

样本均值	样本容量	总体标准差	下限	上限
499.60	80.00	15.00	496.31	502.89

图 4-18 未知总体 σ^2 已知时的均值的 95% 的置信区间

因此,本例所求的未知总体均值的 95% 的置信区间为 [496.31,502.89]。

2. 而当 $D(X)=\sigma^2$ 未知时,在大样本前提下, X 的期望 μ 的 $100(1-\alpha)\%$ 的置信区间为

$$\left[\bar{X} - u_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + u_{\alpha/2} \frac{S}{\sqrt{n}} \right]$$

例 4.18 1973 年某市测量 120 名 12 岁男孩身高资料,得 $\bar{x}=143.05$, $S_e=0.531$,试求该市 12 岁男孩身高均值 95% 的置信区间。

① 在 SPSS 的数据编辑窗口中,建立数据文件 data04-16.sav。其中所建变量及输入的数值见图 4-19。

身高均值	身高标准误
143.05	0.531

图 4-19 data04-16.sav 中的变量及数据值

② 计算下限

按 Transform→Compute Variable 顺序,打开 Compute Variable 对话框(见图 2-53)。在 Target Variable 框中,输入目标变量名为 下限。在 Numeric Expression 框中,输入“身高均值+IDF.NORMAL(0.05/2,0,1)*身高标准误”。

③ 单击 OK 按钮运行,则在数据编辑窗口工作的数据文件中,出现 下限 的新变量及其值。见图 4-20。

身高均值	身高标准误	下限	上限
143.05	0.531	142.01	144.09

图 4-20 未知总体均值的 95% 的置信区间

④ 计算上限

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 *上限*。在 Numeric Expression 框中, 输入 “身高均值+IDF.NORMAL(1-0.05/2,0,1)*身高标准误”。

⑤ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 *上限* 的新变量及其值。见图 4-20。

因此, 本例所求的未知总体某市 12 岁男孩身高均值的 95% 的置信区间为 [142.01, 144.09]。

3. 同理, 当两个总体 X, Y 的分布类型未知, 而总体的方差 σ_1^2 与 σ_2^2 已知时, 要求总体的参数 $\mu_1 - \mu_2$ 的 $100(1-\alpha)\%$ 的置信区间, 可分别从两个总体抽取两个独立的大样本, 根据中心极限定理可知

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

由此仿 4.1.3 中 4 (1) 里的做法可得 $\mu_1 - \mu_2$ 的 $100(1-\alpha)\%$ 的置信区间为

$$\left[(\bar{X} - \bar{Y}) - u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{X} - \bar{Y}) + u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

例 4.19 从体质普查中, 已知山区和湖区的 13 岁男生身高标准差分别为: $\sigma_1 = 6.65 \text{ cm}$, $\sigma_2 = 6.80 \text{ cm}$, 现从山区和湖区的 13 岁男生中分别随机抽取 300 名和 450 名男生进行身高测试, 分别得到 $\bar{x}_1 = 148.2 \text{ cm}$, $\bar{x}_2 = 146.8 \text{ cm}$, 试求山区和湖区的 13 岁男生身高均值差的 95% 的置信区间。

① 在 SPSS 的数据编辑窗口中, 建立数据文件 data04-17.sav。其中所建变量及输入的数值见图 4-21。

山区身高样本 均值	山区身高总体 标准差	山区样本量	湖区身高样本 均值	湖区身高总体 标准差	湖区样本量
148.20	6.65	300.00	146.80	6.80	450.00

图 4-21 data04-17.sav 中的变量及数据值

② 计算下限

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 *下限*。在 Numeric Expression 框中, 输入 “山

区身高样本均值-湖区身高样本均值+IDF.NORMAL (0.05/2,0,1)*sqrt(山区身高总体标准差*山区身高总体标准差/山区样本量+湖区身高总体标准差*湖区身高总体标准差/湖区样本量)”。

③ 单击 **OK** 按钮运行,则在数据编辑窗口工作的数据文件中,出现 下限 的新变量及其值。见图 4-22。

④ 计算上限

按 Transform→Compute Variable 顺序,打开 Compute Variable 对话框(见图 2-53)。在 Target Variable 框中,输入目标变量名为 上限。在 Numeric Expression 框中,输入“山区身高样本均值-湖区身高样本均值+IDF.NORMAL (1-0.05/2,0,1)*sqrt(山区身高总体标准差*山区身高总体标准差/山区样本量+湖区身高总体标准差*湖区身高总体标准差/湖区样本量)”。

⑤ 单击 **OK** 按钮运行,则在数据编辑窗口工作的数据文件中,出现 上限 的新变量及其值。见图 4-22。

山区身高样本均值	山区身高总体标准差	山区样本量	湖区身高样本均值	湖区身高总体标准差	湖区样本量	下限	上限
148.20	6.65	300.00	146.80	6.80	450.00	0.42	2.38

图 4-22 两个未知总体均值差的 95% 的置信区间

因此,本例所求的山区和湖区的 13 岁男生身高均值差的 95% 的置信区间为 [0.42,2.38]。

4. 而当总体的方差 σ_1^2 与 σ_2^2 未知时,在大样本情况下,可用样本方差 S_1^2, S_2^2 来替代总体方差 σ_1^2 与 σ_2^2 ,从而可得 $\mu_1 - \mu_2$ 的 $100(1-\alpha)\%$ 的置信区间为

$$\left[(\bar{X} - \bar{Y}) - u_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{X} - \bar{Y}) + u_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]$$

例 4.20 实测某中学初一男生和初二男生立定跳远,得到 $\bar{x}_1 = 180.5\text{cm}, s_1 = 20.5\text{cm}, n_1 = 160, \bar{x}_2 = 190.5\text{cm}, s_2 = 15.6\text{cm}, n_2 = 158$ 。求初一男生和初二男生立定跳远成绩均值差的 95% 的置信区间。

① 在 SPSS 的数据编辑窗口中,建立数据文件 data04-18.sav。其中所建变量及输入的数值见图 4-23。

初一样本均值	初一样本标准差	初一样本量	初二样本均值	初二样本标准差	初二样本量
180.50	20.50	160.00	190.50	15.60	158.00

图 4-23 data04-18.sav 中的变量及数据值

② 计算下限

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 下限。在 Numeric Expression 框中, 输入 “初一样本均值-初二样本均值+IDF.T(0.05/2,初一样本量+初二样本量-2)*sqrt(初一样本标准差*初一样本标准差/初一样本量+初二样本标准差*初二样本标准差/初二样本量)”。

③ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 下限 的新变量及其值。见图 4-24。

④ 计算上限

按 Transform→Compute Variable 顺序, 打开 Compute Variable 对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 上限。在 Numeric Expression 框中, 输入 “初一样本均值-初二样本均值+IDF.T(1-0.05/2,初一样本量+初二样本量-2)*sqrt(初一样本标准差*初一样本标准差/初一样本量+初二样本标准差*初二样本标准差/初二样本量)”。

⑤ 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 上限 的新变量及其值。见图 4-24。

初一样本均值	初一样本标准差	初一样本量	初二样本均值	初二样本标准差	初二样本量	下限	上限
180.50	20.50	160.00	190.50	15.60	158.00	-14.02	-5.98

图 4-24 两个未知总体均值差的 95% 的置信区间

因此, 本例所求的初一男生和初二男生立定跳远成绩均值差的 95% 的置信区间为 $[-14.02, -5.98]$ 。

4.2.7 估计值的误差限及估计精度

设 X_1, X_2, \dots, X_n 为总体 X 的一个样本, θ 为总体的待估参数, $\hat{\theta}(X_1, X_2, \dots, X_n)$ 为 θ 的一个估计量。对于给定的 $\alpha (0 < \alpha < 1)$, 如果 $P\{|\hat{\theta} - \theta| \leq \Delta(\hat{\theta})\} = 1 - \alpha$, 则称 θ 的置信概率为 $100(1 - \alpha)\%$ 的误差限为 $\Delta(\hat{\theta})$, 相对误差限为 $\Delta'(\hat{\theta}) = \frac{\Delta(\hat{\theta})}{\theta}$, 估计精度为

$$P_c(\hat{\theta}) = 1 - \Delta'(\hat{\theta}) = 1 - \frac{\Delta(\hat{\theta})}{\theta}$$

当抽定一个样本时, 上述定义中的未知参数 θ 可用其估计量 $\hat{\theta}$ 来代替, 因此可求得相应的估计值的误差限、相对误差限及估计精度。

以正态分布为例, 如总体均值 μ 的置信概率为 $100(1 - \alpha)\%$, 则其误差限、相对误差限、估计精度分别为

1. 总体方差 σ^2 已知时

$$\Delta(\bar{X}) = \mu_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\Delta'(\bar{X}) \approx \frac{\mu_{1-\alpha/2} \sigma}{\sqrt{n} \bar{X}}$$

$$P_c(\bar{X}) \approx 1 - \frac{\mu_{1-\alpha/2} \sigma}{\sqrt{n} \bar{X}}$$

2. 总体方差 σ^2 未知时

$$\Delta(\bar{X}) = t_{1-\alpha/2}(n-1) \frac{S}{\sqrt{n}}$$

$$\Delta'(\bar{X}) \approx \frac{t_{1-\alpha/2}(n-1) S}{\sqrt{n} \bar{X}}$$

$$P_c(\bar{X}) \approx 1 - \frac{t_{1-\alpha/2}(n-1) S}{\sqrt{n} \bar{X}}$$

例 4.21 从成年女子总体 X 中随机抽取容量为 100 的样本，测得血红蛋白的 $\bar{x} = 7.36$ ， $s = 0.39$ ，用 95% 的置信区间进行估计时，试求其误差限、相对误差限和估计精度。

① 在 SPSS 的数据编辑窗口中，建立数据文件 data04-19.sav。其中所建变量及输入的数值见图 4-25。

② 计算误差限

样本均值	样本标准差	样本量
7.36	0.39	100.00

图 4-25 data04-19.sav 中的变量及数据值

按 Transform→Compute Variable 顺序，打开 Compute Variable 对话框（见图 2-53）。在 Target Variable 框中，输入目标变量名为误差限。在 Numeric Expression 框中，输入“IDF.T(1-0.05/2,样本量)*样本标准差/sqrt(样本量)”。

③ 单击 OK 按钮运行，则在数据编辑窗口工作的数据文件中，出现 误差限 的新变量及其值。见图 4-26。

④ 计算相对误差限

按 Transform→Compute Variable 顺序，打开 Compute Variable 对话框（见图 2-53）。在 Target Variable 框中，输入目标变量名为 相对误差限。在 Numeric Expression 框中，输入“误差限/样本均值”。

⑤ 单击 OK 按钮运行，则在数据编辑窗口工作的数据文件中，出现相对误差限的新变量及其值。见图 4-26。

⑥ 计算估计精度

按 Transform→Compute Variable 顺序，打开 Compute Variable 对话框（见图 2-53）。在 Target Variable 框中，输入目标变量名为 *估计精度*。在 Numeric Expression 框中，输入“1-相对误差限”。

⑦ 单击 OK 按钮运行，则在数据编辑窗口工作的数据文件中，出现估计精度的新变量及其值。见图 4-26。

样本均值	样本标准差	样本量	误差限	相对误差限	估计精度
7.36	0.39	100.00	0.08	0.01	0.99

图 4-26 计算的误差限、相对误差限和估计精度

因此，本例所求的误差限、相对误差限和估计精度分别为 0.08、0.01 和 0.99。

第 5 章 几种常用的概率抽样方法

5.1 抽 样 概 述

研究对象的全体称**总体**。例如，要研究北京市 13 岁男孩的身高，则所有北京市 13 岁男孩的身高组成研究的总体。显然，总体是由同质的个体所组成的。总体中每个被研究的对象称为**个体**（**总体单元**）。在本例中，北京市每一个 13 岁男孩的身高都可称为个体。总体中所含的个体数称**总体含量**，又称**总体容量**，一般用大写的 N 表示。如果北京市共有 13 岁男孩 16 万人，则本例的总体容量为 16 万。

如果在总体中抽取个体时，每个个体都有相同的概率被抽中，则所抽得的个体的指标 X 是个随机变量，这个 X 的分布就是**总体分布**。称总体分布的分布函数为**理论分布函数**。这个分布函数为：

$F(x)$ =总体中所讨论的指标小于 x 的个体所占的比值。

由于总体中有大量的个体，而每个个体的指标不太可能都能被具体测到，因此，总体分布是客观存在但未知，此时需要借助于从总体中抽取部分个体来对总体分布的某些方面来作推测。这是实际中最常见的做法。一个总体单元（个体）出现在样本中的概率称为**入样概率**。

把上述从总体中抽取的部分个体的集合称为**样本**。样本中所含的个体数称**样本含量**（**容**）**量**（或**样本量**）。样本含量用小写 n 表示。样本从总体中抽取的方式称**抽样**。显然，一个样本对总体要有好的代表性必须满足抽样是随机的，使总体中每个个体都有同等的机会被选入到样本中去，而且样本含量必须足够，也即要达到研究设计中所提出的研究精度的最低要求。在这种条件被满足时，当样本含量较大时，根据格列汶科定理，样本的经验分布与总体分布是一致的。

因此，可用样本的分布情况去推断总体的分布情况。

如果研究中每个个体入样概率已知且抽样方式又是随机的，这种抽样统称为**概率抽样**。它常用于定量研究或大规模研究中。

在实际的抽样调查中，由于涉及范围广、问题相对复杂，加上各被调单位的规模有很大差异，导致总体中的各个单元并不都处在同等的地位，因此对各单元一概采用简单随机抽样的方式并不能保证一定得到对总体具有很好代表性的样本，且也不一定最经济。

因此，需要通过对所研究的总体确定是否要进行分层、分群，再根据实际研究精度、在抽样阶段中合理地选用下述部分或全部的抽样方法，对获取一个完整的样本过程，进行周全的设计，这样得到的方案称为抽样设计方案。一般地，抽样设计方案由如何抽样和根据选定的抽样方法进行估值两个基本部分组成。

在 SPSS 的复合抽样程序中，可以进行简单的概率抽样，也可以进行多种概率抽样方法叠加在一起所组成的复合抽样，以及根据这些设计的抽样方法来对总体做出相应的估值。在 SPSS 中，提供了九种可供选择的抽样方法，不同的抽样方法对应于不同的估值方法，考虑到篇幅有限，我们只选取其中最为常用的四种抽样方法给以介绍，它们分别是：

- ① Simple Random Sampling（简单随机抽样法）。
- ② Simple Systematic（简单系统抽样法）。
- ③ PPS（PPS 抽样法）。
- ④ PPS Brewer（PPS Brewer 抽样法）。

另外，介绍几个在大规模抽样调查中可能用得着的、建立在这些抽样方法基础上的分层随机抽样、整群随机抽样和多阶随机抽样。

5.2 简单随机抽样

当总体只有总体各个单元的名录，而没有哪些个体是更加重要的辅助信息，只能对它们一视同仁时，也即总体仅有一个简单的抽样框时，我们通常采用简单随机抽样。

所谓**抽样框**是指对可以选择作为样本的总体单位列出名册或排序编号，以确定总体的抽样范围和结构。它又称“**抽样框架**”、“**抽样结构**”。设计出了抽样框后，便可采用抽签的方式或按照随机数表来抽选必要的单位数。若没有抽样框，则不能计算样本单位的概率，从而也就无法进行概率选样。一个好的抽样框应做到完整而不重复。大学学生花名册、城市黄页里的电话列表、工商企业名录、街道派出所里居民户籍册、意向购房人信息册等都是常见的抽样框。例如：要从 10000 名职工中抽出 200 名组成一个样本，则 10000 名职工的名册，就是抽样框。在没有现成的名单的情况下，可由调查人员自己编制。应该注意的是，在利用现有的名单作为抽样框时，要先对该名录进行检查，避免有重复、遗漏的情况发生。以提高样本对总体的代表性。

简单随机抽样是在总体的 N 个有限个体（抽样单元）中取 n 个个体组成样本时，使每个样本出现的概率均等于 $1/C_N^n$ 的一种基本的抽样方法。这是最基本的抽样方式，称为简单随机抽样。它是其他抽样方法的基础。其他抽样方法可以看作是对它的修正。

5.2.1 样本容量的确定

5.2.1.1 按绝对精度决定样本含量

1. 计算公式

在给定绝对精度 d 和 $1-\alpha$ 的置信度时, 要求 $|\bar{y} - \bar{Y}| \leq d$, 即

$$P\{|\bar{y} - \bar{Y}| \leq d\} = 1 - \alpha,$$

式中, \bar{y} 是样本均值; \bar{Y} 是总体均值。

根据正态分布区间估计的理论可得: $d = \mu_{1-\frac{\alpha}{2}} \sqrt{\nu(\bar{y})}$,

式中 $\nu(\bar{y})$ 是样本方差的期望, $\mu_{1-\frac{\alpha}{2}}$ 是 $1-\alpha/2$ 处的标准正态分布的位置值。

所以

$$d^2 = (\mu_{1-\frac{\alpha}{2}})^2 \nu(\bar{y}) = (\mu_{1-\frac{\alpha}{2}})^2 \frac{1}{n} (1 - \frac{n}{N}) S^2$$

由此可得

$$n = \frac{(\mu_{1-\frac{\alpha}{2}})^2 S^2}{d^2 + \frac{1}{N} (\mu_{1-\frac{\alpha}{2}})^2 S^2}$$

当总体容量 N 很大时, $n \approx \frac{(\mu_{1-\frac{\alpha}{2}})^2 S^2}{d^2}$

由上式可知, 当总体容量 N 很大时, 样本含量 n 与总体含量 N 关系不是很大, 此时, 关键取决于总体的方差。

2. 实例分析

例 5.1 某社区有居民 300 户, 共 1100 人, 为抽样调查该社区居民每月每户用以食物的消费支出, 要求平均每月每户用以食物的消费支出的估计绝对误差不超过 40 元, 应调查多少户?

【题析】 本例已知 $d = 40$, 但总体 S^2 未知, 需要首先获取 S^2 的一个粗略估计值。常用的方法有三种。

(1) 查阅资料法

查阅有关资料，如果所研究的总体以前被调查过，则可以用以前调查获得的 S^2 计值作为粗略估计值。

(2) 预先调查法

对总体先随机抽取一个样本含量较小的样本，用该样本得到的方差作为总体方差 S^2 的估计值，再用这个粗略估计值去确定所需的样本含量，当这个确定的样本含量大于预先抽取的样本含量时，只需再补充调查不足部分的样本单元使之达到所需的样本含量即可。

(3) 类推法

通过查阅有关资料，如果能找到与目标量 Y 高度关联的指标量 X 的信息，当 \bar{X} 和 S_X^2 有估计值，且 X 与 Y 的变异系数接近，则预先抽取一个样本含量较小的样本则很容易得到 \bar{Y} 的估计值，根据 $S_Y / \bar{Y} \approx S_X / \bar{X}$ ，容易得到 S^2 的粗略值。

本例在没有先前资料可查，也无法用类推的方法获取 S^2 的一个粗略估计值时，我们首先用预先调查法来实现样本含量的估计。

具体步骤如下：

第一步：先用简单随机抽样法，在 300 户中随机抽取 35 户（这个户数是预估数，抽 40 户、30 户或其他较小的值均可）。

具体做法参见 5.2.2 小节中的做法。

抽样框的资料及抽中的 35 户的入样概率、抽样权重、总体容量等相关资料已存放在数据文件 data05-01a.sav 中。该数据文件中的英文名称分例 5.4 在 SPSS 中进行简单随机抽样后由 SPSS 自动生成，对其中文解释参见 215 页。

第二步：对 35 户随机抽中的居民进行调查，收集包括户人数、人均月收入 and 户月食物支出三个指标在内的样本数据资料，见表 5-1。该资料已经存放在数据文件 data05-02.sav 中。

表 5-1 35 户样本数据资料

	编号	户人数	人均月收入	户月食物支出
1	2	2	710.00	540.00
2	13	3	557.00	830.00
3	14	4	408.00	860.00
4	24	4	458.00	1070.00
5	27	3	440.00	710.00
6	32	5	410.00	1100.00
7	41	2	825.00	610.00
8	56	2	730.00	660.00
9	58	4	470.00	1010.00
10	60	5	356.00	1030.00
11	83	3	640.00	880.00
12	92	4	570.00	1040.00
13	103	3	590.00	790.00
14	110	2	940.00	640.00
15	111	4	423.00	960.00
16	124	4	498.00	1100.00
17	134	3	550.00	770.00
18	137	4	663.00	1270.00
19	142	5	524.00	1440.00
20	146	3	630.00	890.00
21	148	3	610.00	730.00
22	156	6	325.00	1230.00
23	160	3	873.00	1090.00
24	162	3	607.00	790.00
25	182	2	885.00	590.00
26	190	4	400.00	900.00
27	198	2	345.00	620.00
28	207	4	488.00	970.00
29	219	4	423.00	980.00
30	232	2	950.00	660.00
31	234	4	418.00	1000.00
32	260	5	370.00	1170.00
33	268	2	815.00	600.00
34	278	3	563.00	640.00
35	287	4	475.00	980.00

第三步：计算样本方差，用样本方差作为总体方差的粗略估计值。

按 Analyze→Descriptive Statistic→Descriptives 顺序，展开 Descriptives 对话框，从其左侧变量名源框中，选择户月食物支出变量，移入到 Variable[s] 下框中，单击 OK 按钮，则在输出窗中，出现输出结果，见表 5-2。

表 5-2 描述统计结果

Descriptive statistics

	N	Mean	Std Deviation	Variance
户月食物支出	35	8.9571E2	217.75979	4.742E4
Valid (listwise)	35			

由此可得总体方差 S^2 的粗略估计值为 4.742×10^4 。

第四步：计算所需绝对误差限下所需的样本容量。

① 按 File→New→Data 顺序，打开一个新的数据编辑窗口。建立一个数据文件 data05-03.sav, 其中变量名为：总体容量、绝对误差限和方差，它们的观测值分别为：300、40 和 47420。

② 在 Transform→Compute Variable 顺序，展开 Compute Variable 对话框。

③ 在 Target Variable 下框中输入样本含量，在 Numeric Expression 下框中输入以下双引号中的表达式：

“IDF.NORMAL(0.975,0,1)*IDF.NORMAL(0.975,0,1)*方差/(绝对误差限*绝对误差限+1/总体容量*IDF.NORMAL(0.975,0,1)*IDF.NORMAL(0.975,0,1)*方差)”

函数中，第一个参数 0.975 由 $1-0.05/2$ 计算得到，第二和第三个参数 0, 1 为标准正态分布的均值和方差。

④ 单击 OK 按钮，在数据文件中增加一个新变量：样本含量，以及其计算结果的值 82.53。

这说明，要满足绝对误差限的要求，至少要调查 83 户。一般地，由于这是概算结果，所以，在实际中为保证研究的精度，一般应在此基础上增加 10%左右的样本量，也即本例应调查 91 户左右。在实际调查中，只需在 35 户基础上再增补调查 56 户即可。

例 5.2 为调查北京市家庭汽车普及情况，拟在北京市进行一次家庭抽样调查，要求所估计的百分数误差不超过 2%，则要抽取一个简单随机样本，应取多大样本容量？

【题析】 由于北京市家庭户数达到数百万，又已知 d 值为 2%，所以，测算样本含

量时，可以用 $n \approx \frac{(\mu_{1-\frac{\alpha}{2}})^2 S^2}{d^2}$ 来进行计算。因此，关键要知道 S^2 的一个粗略值，即可。

设北京市家庭汽车普及率为 P , $P = \frac{N_1}{N}$, N_1 为北京市拥有至少一辆汽车的家庭总数。

N 为北京市家庭总数。则根据总体方差的定义可知： $S^2 \approx P(1-P)$ 。

按最保险的情况估计样本含量时, $S^2 \approx P(1-P)$ 应取最大值, 显然此时的 $P = 0.5$, 故 $S^2 \approx P(1-P) = 0.5 \times 0.5 = 0.25$ 。

根据以上分析, 现可以计算样本含量了, 在 SPSS 中的具体步骤如下:

① 按 File→New 顺序, 打开一个新的数据编辑窗口。建立一个数据文件 data05-04.sav, 其中变量名为: 绝对误差限和最大方差, 它们的观测值分别为: 0.02 和 0.25。

② 按 Transform→Compute Variable 顺序, 展开 Compute Variable 对话框。

③ 在 Target Variable 下框中输入 *样本含量*, 在 Numeric Expression 下框中输入以下双引号中的表达式:

“IDF.NORMAL(0.975,0,1)*IDF.NORMAL(0.975,0,1)*最大方差/(绝对误差限*绝对误差限)”。

函数中, 第一个参数 0.975 由 $1-0.05/2$ 计算得到, 第二和第三个参数 0, 1 为标准正态分布的均值和方差。

④ 单击 OK 按钮, 在数据文件中增加一个新变量: 样本含量, 以及其计算结果的值 2400.91。

这说明, 要满足绝对误差限 2% 的要求和达到最保险的情况, 至少要调查 2401 户。

5.2.1.2 按相对精度决定样本含量

1. 计算公式

在给定相对精度 h 和 $1-\alpha$ 的置信度时, 要求 $\left| \frac{\bar{y} - \bar{Y}}{\bar{Y}} \right| \leq h$, 即

$$P\left\{\left|\frac{\bar{y} - \bar{Y}}{\bar{Y}}\right| \leq h\right\} = 1 - \alpha$$

根据正态分布的区间估计理论可得: $\bar{Y}h = \mu_{1-\frac{\alpha}{2}} \sqrt{v(\bar{y})}$

$$\text{由此可得: } n = \frac{(\mu_{1-\frac{\alpha}{2}})^2 S^2}{(\bar{Y}h)^2 + \frac{1}{N}(\mu_{1-\frac{\alpha}{2}})^2 S^2} = \frac{(\mu_{1-\frac{\alpha}{2}})^2 C^2}{h^2 + \frac{1}{N}(\mu_{1-\frac{\alpha}{2}})^2 C^2}$$

其中 $C = S/\bar{Y}$ 为变异系数, 当总体容量 N 很大时, 可取

$$n \approx \frac{(\mu_{1-\frac{\alpha}{2}})^2 C^2}{h^2}$$

2. 实例分析

例 5.3 为调查目前北京市家庭汽车普及情况, 拟在北京市进行一次家庭抽样调查, 要求所估计的相对误差不超过 10%, 前两年的抽样调查表明北京市的家庭汽车普及率为 15%-25%, 则要抽取一个简单随机样本, 应取多大样本容量?

【题析】 由于北京市家庭户数达到数百万, 所以, 测算样本含量时, 可以用

$$n \approx \frac{\left(\mu_{1-\frac{\alpha}{2}}\right)^2 C^2}{h^2}$$

来进行计算。由于已知估计的相对误差 h 为 10% 因此, 关键要知道 C^2 的一个粗略值, 即可。

在假设北京市家庭汽车普及率为 P 的前提下, 变异系数 $C = \sqrt{(1-P)/P}$ 。

由变异系数的计算公式可知, 当 P 从 0 到 1 增大时, $1-P$ 的值则减少, 当 P 取最小值时, C 取得最大值。根据前两年的抽样调查结果, 可以肯定现在北京市的家庭汽车普及率不会低于 15%。所以, 最大的 $C^2 = (1-0.15)^2 / 0.15 = 0.85^2 / 0.15$ 。

根据以上分析, 现可以计算样本含量了, 在 SPSS 中的具体步骤如下:

① 按 File→New 顺序, 打开一个新的数据编辑窗口。建立一个数据文件 data05-05.sav, 其中变量名为: 相对误差, 其观测值为: 0.1。

② 按 Transform→Compute Variable 顺序, 展开 Compute Variable 对话框。

③ 在 Target Variable 下框中输入 *样本含量*, 在 Numeric Expression 下框中输入以下双引号中的表达式:

“IDF.NORMAL(0.975,0,1)*IDF.NORMAL(0.975,0,1)*0.85*0.85/0.15/(相对误差*相对误差)”

④ 单击 OK, 在数据文件中增加一个新变量: 样本含量, 以及其计算结果的值 1850.30。

这说明, 要满足相对误差 10% 的要求和达到最保险的情况, 至少要调查 1851 户。

5.2.2 简单随机抽样过程

1. 简单随机抽样的基本做法:

为了从一个总体中选择一个简单随机样本, 我们需要知道总体中所有研究对象的一个名单。称这个名单为 **抽样框 (sampling frame)**。假设你打算对北京市的家庭做一个抽样

调查, 总体是北京市所有家庭。一个可能的抽样框是从公安局获取的各家庭的户籍名目。

选择一个简单随机样本的极大多数常用的方法是: ①在抽样框中给研究对象编号, ②随机地产生这些编号的一个集合, 以及③抽样调查那些产生编号的研究对象。要求使用随机数去选择样本以确保每个被试对象有同等挑选的机会。

2. 在 SPSS 中实现简单随机抽样:

例 5.4 对例 5.1 中某社区 300 户居民用简单随机抽样的方法, 抽取一个 35 户的随机样本。

步骤如下:

① 在 SPSS 中数据编辑窗中, 先建立如图 5-1 所示的 300 户抽样框数据文件, 见 data05-01.sav。

② 单击 Analyze→Complex Samples→Select a Sample 顺序, 打开 Welcome to Sampling Wizard 对话框, 见图 5-2。

图 5-1 300 户编号



图 5-2 Welcome to Sampling Wizard 对话框

Sampling Wizard 提供了创建、修改或执行一个抽样设计方案文件的完整的步骤。在使用它之前, 应在头脑中首先有一个做过定义的目标总体、一张抽样个体的名录和一个适当的抽样设计方案。其次, 要在 Sampling Wizard 对话框中回答: What would you like to do?

答案可从以下三个选项中分别加以选择:

- (1) Design a sample。需要建立一个抽样方案文件时, 可选择本选项。
- (2) Edit a sample design。已经有了一个抽样设计方案, 现要对已建立的该方案进行修改、完善时, 可选择本选项。
- (3) Draw the sample。用已建立的抽样设计方案文件去抽取一个样本时, 可选择本选项。

本例由于从一无所有开始做起, 所以, 在 Sampling Wizard 对话框中, 选择 Design a sample 选项, 将插入点定位在其后 File 的文本编辑框中, 通过直接输入 “E: \第 5 章 抽样方法\data5-01.csplan” 来定义抽样方案文件名。它也可以通过单击 Browse 按钮, 在弹出的选项卡中, 通过逐级选择存放路径和最后定义文件名的方式来完成, 具体做法参见

第2章中的相关内容。

3. 单击 Next 按钮, 进入 Stage 1: Design Variables 对话框, 见图 5-3。Design Variables 对话框有三部分主要内容组成:

第一部分: 最左侧的矩形框中, 列出了由抽样设计中各步骤的要点组成的树状目录结构控制系统, 它可以用来操控 Sampling Wizard。

在对抽样设计方案所有细节的定义过程中, 一般都需要通过很多步骤才能完成。所以, 每一步

Sampling Wizard 的左侧矩形框中都会列出所有步骤的要点, 你可以通过单击在要点中激活步骤的名字来操纵 Wizard。如果每个以前的步骤对该步骤已经给出了最小需求说明, 则它是有效的。只要所有先前步骤是有效的, 则步骤就是激活的。激活步骤的字体颜色变为黑色。并非所有步骤的要点都是有效的, 对于为什么一个给定步骤可能是无效方面的更多资料可参阅个别步骤的 Help。

第二部分: 正中间 Variables 下面的矩形框中, 列出了当前数据文件中的变量名列表; 可用它对右侧第三部分的内容作相应的定义。

第三部分: 最右边给出的是需要对抽样设计方案中进行定义的各细节部分。它对应于在第一部分中所要做出的选择。

在最初进入 Design Variables 对话框时, 第一阶的 Design Variables、Sampling Method 和 Sample size 处在激活状态, 只要单击选中的步骤名称即可展开其对话框。这意味着你可以不分先后地对它们做出定义, 也即表明在已进入其他步骤后, 单击左侧矩形框中的 Design Variables, 也可进入 Design Variables 对话框。下同。

在 Design Variables 对话框中, 最右侧部分是用来设计变量的。它最多共有 4 个部分需要使用者来定义。从上到下它们分别是层和整群变量、抽样权重, 以及为这一步指定一个标签。

(1) 设定层变量

层是指将总体所划分成的若干个不相交的 k 个子总体。一般各层有大体一致的数量较少的单元个体。交叉分组的层变量定义了独立的子总体或层。获取各层各自的样本。为提高估计的精度, 作为重要的特征, 各层内的单元应尽可能同类。

在 Variables 下面的源变量名框中选择一个变量, 单击右移箭头将其移入到 Stratify By 的下框中, 完成层变量定义。

(2) 设定群变量

群是层中所划分的小总体。

在 Variables 下面的源变量名框中选择一个变量, 单击右移箭头将其移入到 Clusters

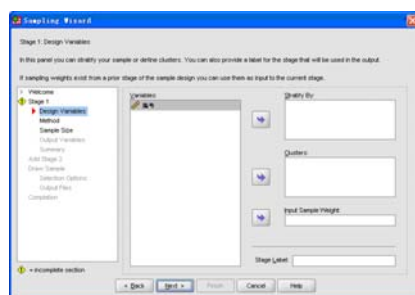


图 5-3 Design Variables 对话框

(3) 设定抽样权重变量

在 Variables 下面的源变量名框中选择一个变量，单击右移箭头将其移入到 Input Weight 的下框中，完成抽样权重变量定义。

在 Stage Label 的后框中，可以输入任意的字符串标签，以便在输出里用它来帮助识别分阶段的信息。它是任选项，可以不选。

4. 单击 Next 按钮，进入 Stage 1: Sampling Method 对话框，见图 5-4。在 Sampling Method 对话框中，可以定义抽样方法和规模测度。这可以详细说明如何从工作的数据集选样品。

① 确定抽样方法种类

- Simple Random Sampling (简单随机抽样法)。等概选择单元。可用有放回或无放回的方法选择它们。

-

图 5-4 Sampling Method 对话框

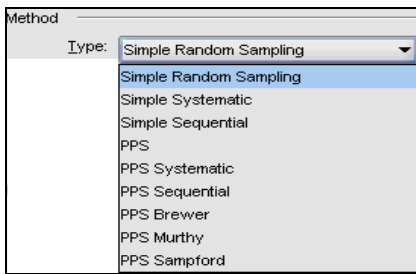


图 5-5 选择抽样类型

- **Simple Sequential** (简单连续抽样法)。用等概和无放回抽样的方法连续地选择单元。
- **PPS** (PPS 抽样法)。这是第一阶段中使用的方法, 它用抽取概率正比于规模测度抽样方法随机选择单元。任何单元都可以重复地选择, 只有整群可以无放回地抽样。
- **PPS Systematic** (PPS 系统抽样法)。这是第一阶段中使用的方法, 用概率正比于规模测度抽样方法系统地选择单元。单元是无放回抽取的。
- **PPS Sequential** (PPS 连续抽样法)。这是第一阶段中使用的方法, 用抽样概率正比于整群规模测度抽样方法连续且无放回地选择单元。
- **PPS Brewer** (PPS Brewer 抽样法)。这是第一阶段中使用的方法, 用抽样概率正比于整群规模测度抽样方法从每个层中无放回地选择两个整群。使用本方法必须指定一个整群变量。
- **PPS Murthy** (PPS Murthy 抽样法)。这是第一阶段中使用的方法, 用抽样概率正比于整群规模测度抽样方法从每个层中无放回地选择两个整群。使用本方法必须指定一个整群变量。
- **PPS Sampford** (PPS 浅层抽样法)。这是第一阶段中使用的方法, 用抽样概率正比于整群规模测度抽样方法从每个层中无放回地选择两个以上的整群。它是 **Brewer's method** 的推广。使用本方法必须指定一个整群变量。

如果你选择 **PPS Brewer** 或 **PPS Murthy** 抽样方法, 则你可以单击 **Finish** 按钮, 抽取一个样本。

值得一提的是, 某些 **PPS** 的抽样方法只有当已定义群变量时才有效, 所有 **PPS** 的抽样方法在设计的第一阶段中都有效, **WR** 方法只在设计的最后一阶段中有效。

② 选择抽样方式

有以下两种方式可供选择:

- **Without replacement (WOR)** (无放回)。在抽样过程中, 从总体中抽取的个体, 不再放回原总体。选择任一种抽样方法, 都可选择本选项。它是系统默认选项。
- **With replacement (WR)** (有放回)。在抽样过程中, 从总体中抽取的个体, 记录后再放回原总体, 继续参与后续的抽样。需要注意的是: 只有选择 **Simple Random Sampling**、**PPS** 抽取单元时, 可以选择本选项。

抽样方式的不同, 它决定了计算抽样概率的方法是不一样的, 它会影响到后续的估值方法的不同。

③ 选择分析中的估计方法

在默认情况下, 在方案文件里指定的估计方法同选择的抽样方法是一致的。即使抽样方法暗示用 **WOR** 估计, 但你可以通过选择 **Use WR estimation for analysis** 选项, 来改变原先暗示的 **WOR** 估计, 而使用 **WR** 估计。本选项只在抽样的第一阶段中有效。

无放回的不等概抽样比有放回的不等概抽样, 原则上可获得较多的信息, 有更高的

效率，但在抽取样本与样本数据的分析计算等方面无放回抽样均要比有放回抽样复杂些。

当抽样比 $f = \frac{n}{N}$ 很小时，无放回抽样比有放回抽样提高的效率有限。

(2) 设定规模测度

在 Measure of Size (MOS) 栏中，规模测度可以明确地用一个变量定义，或可以从数据中计算得到。任意地，在 MOS 中，你可以设置下界和上界，它们优先于建立在 MOS 变量中的值或从数据中计算得到的值。这些任选项只在第一阶中可用。

需要注意的是，在 Method Type 中选择前三种抽样方法时，本选择项是无效的。只有当选择一个 PPS 方法时，规模测度选项才被激活。PPS 方法一般用以整群抽样。

① Read from variable 选项，如果在当前工作的数据文件的一个变量中已明确地定义了各个单元容量，选择本选项，并从 Variable 选择该变量，将其移入到本选项的下框中。

② Count data records 选项，如果各个单元的容量需从数据中计算得到，则选择本选项。在 Minimum 后框中输入容量的最小值，在 Maximum 后框中输入容量的最大值。以此来任意地设置 MOS 的下限和上限。

本例不作任何选择，即采用系统默认的选择 simple random sampling。同时，采用系统默认的 without replacement(WOR)选项，即在本阶段中，利用默认的不重复（WOR）简单随机抽样法（simple random sampling），抽取的编号作为主要的抽样单元。

5. 单击 Next 按钮或在左侧矩形框中单击 Sample Size，可打开 Sample Size 对话框，见图 5-6。在本对话框中可以设定样本容量。

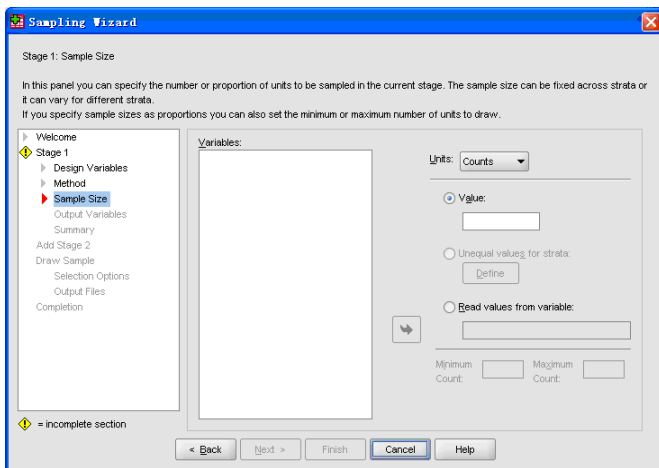


图 5-6 Sample Size 对话框

在多阶抽样中，为了指定样本容量，可以用在前面阶段里选取的整群变量来作为本阶段的分层变量。

(1) 确定抽样单位

在 Units 下拉列表中, 有两项可供选择:

- ① Counts 选项, 如果要为样本指定一个精确的样本容量, 则选择本选项。
- ② Proportions 选项, 如果要选取单元的比例, 则选择本选项。

(2) 输入抽样的数量或比例

在以下三种方式中任意选择一种, 来完成对抽样数量的定义。

① Value 选项, 在这里输入的值应用于所有的层。如果在上面已选择 Counts 作为单位量纲, 则应该输入一个正整数。如果在上面已选择比例作为单位量纲, 则应该输入一个正数值。除非是在上面选择了有放回抽样, 否则, 比例值也应小于等于 1。由于总体抽样单元的数量与比例的乘积不一定是正整数, 所以实际抽取的样本单元的比例可以大于等于在这里输入的比例。

② Unequal values from variable 选项, 如果各层选取的容量不同, 选择本选择项。然后单击 Define 按钮, 展开 Define Unequal Sizes 对话框, 见图 5-7。在 Define Unequal Sizes 对话框中, 可以在每个层里输入容量值。



图 5-7 Define Unequal Sizes 对话框

Size Specifications grid (容量说明格)。容量说明格显示了层或群变量五个的交叉分组——每行是一个层/群的组合。符合条件的容量说明格变量包含当前和先前阶段的全部的层变量和先前阶段的全部的群变量。单击层变量或群变量名称, 单击右键, 用弹出的快捷菜单可对格中的变量进行重新排序或将选中的层或群变量拖曳或单击中间的右移箭头将其移到 Exclude 列表。在最右边列中输入相应的容量值。

单击 Labels 或 Values 按钮, 可为容量说明格单元中的层和群变量切换显示值标签和数据值。无标签值的单元总是显示值。

单击 Refresh Strata (更新层) 按钮, 可为容量说明格变量用各个标签数据值组合重新赋值到容量说明格中。

Exclude (排除)。为了给层/群的子集组合指定容量, 移动一个或多个变量到 Exclude 列表。表明不用给这些变量来定义样本容量。

完成设置后, 单击 Continue 按钮, 返回图 5-6 样本容量选择项对话框。

③ Read values from variable 选项, 如果在数据文件中已定义了一个包含各层(群)容量值的变量, 可选择本选项。此时, 可以从变量列表选择一个包含各层(群)容量值的数字型变量并将它移到本选项的后框中。

如果选择的是包含容量比例的变量, 则还要在 Minimum count 和 Maximum count 中去设置抽取单元的数量的上、下限。

在本例中, 在 Units 下拉列表中, 选择 Counts, 因此, 在 Values 下框中输入 35。表

示在本阶段中选择 35 个样本单元。

6. 单击 Next 按钮, 进入 Stage 1 Output Variable 对话框, 见图 5-8。在本步骤中, 当抽取一个样本时, 你可以为存储的样本选取存入的变量。

Which variables do you want to save? (你想要存储那些变量?) 问题中共有四个选择项, 可以选取其中一项或多项。

(1) Population size (总体容量)。估计一个特定阶段中总体中的单元数。存储变量的根名为 PopulationSize_。

(2) Sample proportion (抽样比例)。在特定的阶段里的抽样比例。存储变量的根名为 SamplingRate_。

(3) Sample size (样本容量)。在特定的阶段里抽样单元的数量。存储变量的根名为 SampleSize_。

(4) Sample weight (样本权重)。它等于入样概率的倒数。存储变量的根名为 SampleWeight_。

此外, 在存储的数据文件中, 还自动地产生一些阶段抽样中形成的新变量。它们包括:

Inclusion probabilities (入样概率)。在特定的阶段里抽取单元的比例。新变量的根名为 InclusionProbability_。

Cumulative weight (累积权重)。在前面所有阶段到当前这一阶段里累积抽样权重。新变量的根名为 SampleWeightCumulative_。

Index (索引)。标识特定的阶段内多次选择的单元。新变量的根名为 Index_。

注解: 新变量的根名包括一个反映阶段数的整数后缀。例如: PopulationSize_1_存储第一阶段里的总体容量。

本例在输出变量选项卡中选择全部四项, 要求在数据文件中存取总体容量、样本容量、样本比例和样本权重变量。

7. 单击 Next 按钮, 进入 Stage 1: Plan Summary 对话框, 见图 5-9。查看一下抽样设计。

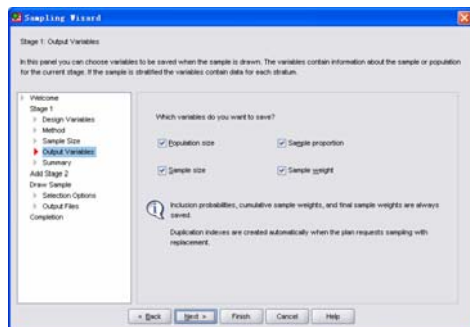


图 5-8 Output Variable 对话框

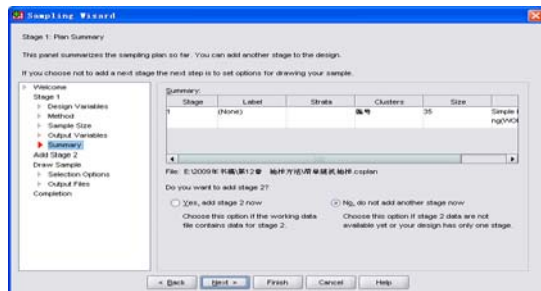


图 5-9 Plan Summary 对话框

这是每个阶段的最后一步，提供直到现阶段为止的抽样设计说明的摘要。从此，既可以继续进入下一阶段（如果需要），也可以直接用设置的选项抽取样本。

左侧矩形框列出了激活的步骤名称，红三角指向当前的步骤。

在 **Summary** 中，详细地列出了已经建立的抽样方案。你可以从中重新审视一下已经建立的抽样方案。

在其下面，显示抽样设计方案的文件名，及其存放的位置。

（1）决定去向

回答问题 “Do you want to add stage 2?”，有两个决定后面去向的选项：

① Yes, add stage 2 now 选项，如果要继续对方案输入更多的要求，选择本选项。

② No, do not add another stage now 选项，如果你不想用其他阶段向方案中增加其他的信息了，则选择本项。它是系统默认选项。

需要指出的是，如果选择 Yes，则进入第二阶段，重复上述 2-6 中的一个步骤或多个甚至全部步骤。如果需要，还可进入第三个阶段进行抽样方案的设计。基本操作方法同第一阶段的做法。其余类推。

（2）修改

单击 **Back** 按钮或单击左侧矩形框中上面的各项可对所作的方案进行及时的更改。

（3）执行

单击 **Finish** 按钮，则递交执行。在输出窗中，出现你所设计的方案说明。在原始数据集中出现选中的单元及相关说明。

在本例中，保持系统默认选择，不再增加其他阶的抽样。

8. 单击 **Next** 按钮，进入到 **Stage 2: Selection Options** 对话框，见图 5-10。在本对话框中，有两个问题和两个任选项需要应答和选择。

（1）决定是否抽样

在 **Do you want to draw a sample** 下，有 Yes 和 No 两个备选项。对它的回答决定了是否要马上做抽样工作。

① Yes 选项，系统默认选择项。选择该项，意味着要抽取一个样本。可在 **Stages** 的下拉式列表中选择要抽取哪个阶段的样本。你可以选择执行部分抽样设计，但必须单击阶段的顺序抽样，即除非已抽取第一阶段的样本后，才可以对第二阶段进行抽样。在编辑或执行一个方案时，你不能对已完成抽样的阶段重新抽样。

② No 选项，选择本选择项，表示还不想抽取一个样本。

（2）决定使用何种种子值

What type of seed value do you want to use 下，有 **A randomly-chosen number** 和 **Custom value** 两个备选项。对其回答决定了抽样中使用的种子值的产生的方式。

① **A randomly-chosen number** 选项，选择该项，由系统随机选择一个数作为产生随

机数的种子值。该项为系统默认选项。

② Custom value 选项, 选择该项, 由用户自定义一个作为产生随机数的种子值, 并在其随后的方框中输入该值。

(3) 决定用户缺失值是否有效

Include in the sample frame cases with user-missing values of stratification or clustering variable 选项, 它决定用户的缺失值是否是有效的。如果是, 选择本项, 则用户的缺失值会作为单独的一类处理。

(4) 决定工作数据是否事先已做过按类排序

Working data are sorted by stratification variables (presorted data may speed processing) 选项, 如果抽样框已用分层变量值预先排序分类, 则选择本项可加速选择过程。

(5) 修改

单击 Back 按钮或单击左侧矩形框中上面的各项可对所作的方案进行及时的更改。

(6) 执行

单击 Finish 按钮, 则递交执行。在输出窗中, 出现你所设计的方案说明。在原始数据集中出现选中的个体及相关说明。

在本例中, 在 What type of seed value do you want to use 下, 选择 Custom Value, 输入 241972 作为随机种子值。其他采用系统默认选项。

由于采用的是随机抽样, 所以, 即使输入相同的种子值, 再次运行后, 采样结果也不一定相同。

9. 单击 Next 按钮, 进入到 Stage 2: Output Files 对话框, 见图 5-11。在对话框中, 可以去选择样本数据、联合概率和样品选择规则存储的地点。

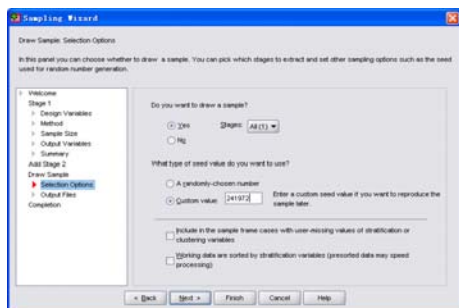


图 5-10 Selection Options 对话框

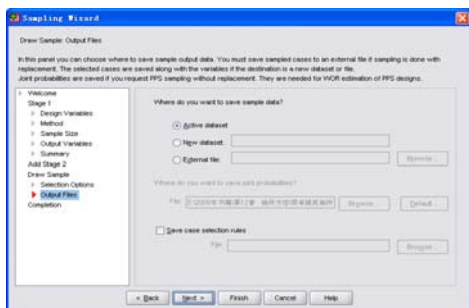


图 5-11 Output Files 对话框

(1) 选择存储样本数据的位置

在 Where do you want to save sample data? 下, 决定输出样本存储在何处。它有三个备选应答项:

① **Active dataset** 选项, 选择本选项, 是将选择样品的抽样输出变量增加到工作数据集里。

② **New dataset** 选项, 选择本选项, 是将选择样品的抽样输出变量和工作数据集里变量存储到新的数据集中。同时, 在 **New dataset** 后框中, 你需要给出新的数据集名称。

③ **External file** 选项, 选择本选项, 是将选择样品的抽样输出变量和工作数据集里变量存储到一个外部的 **SPSS** 格式的数据文件里。同样, 在 **Files** 后的方框中输入外部文件名, 单击 **Browse** 按钮, 为外部文件名选取存放的盘符和子目录。

在当前版本中数据集是有效的, 但在以后的版本中除非你明确地把它们作为数据文件存储, 否则数据集是无效的。数据集的命名必须坚持变量命名规则。

(2) 选择存储联合概率的地方

Where do you want to save joint probabilities 选择项, 决定联合概率存储在何处。它们存储到一个外部的 **SPSS** 格式数据文件中。

它提供了两种存储文件的途径。

① **Default** 选项, 选择本选项, 可将文件存储在原文件中。

② **file** 选项, 选择本选项, 可自定义文件名, 在其后的方框中输入路径、文件名, 或单击 **Browse** 按钮, 为外部文件名选取存放的盘符和子目录。

如果选择的是 **PPS WOR**、**PPS Brewer**、**PPS Sampford** 或 **PSS Murthy** 方法且不指定 **WR** 估计, 则产生联合概率。

(3) 存储样品选择规则

Save case selection rules 选项, 如果你一次构建抽样的一个阶段, 则需要将样品选择规则存储到一个文本文件中。它们为后续阶段创建子抽样框是有用的。此时应选择本选项。

(4) 修改

单击 **Back** 按钮或在左侧矩形框中单击上面的某个要点项, 可对该步所作的方案进行及时的更改。

(5) 执行

如果全部设置完成, 则可单击 **Finish** 按钮, 递交执行。在输出窗中, 出现你所设计的方案说明。在原始数据集中出现选中的个体及相关说明。

在本例中, 保持系统默认选项, 即将样本数据存在工作的数据文件中。

10. 单击 **Next** 按钮, 进入到 **Stage 2: Completion** 对话框, 见图 5-12。这是最后一步。你可以存储方案文件并立即抽取样本, 或粘贴你的选择到语句窗口中。

(1) 回答你要做什么

在 **Completion** 选项卡中, 只有一个问题 **What do you want to do** 需要回答。

它有两个应答项:

① Save the design to a plan file and draw the sample 选项，如果你想要编辑的方案覆盖已存在的方案文件或把新的设计存储到已存在的方案文件并抽取样本，则可选择本选项。

② Paste the syntax generated by the Wizard into a syntax window 选项，如果你想要把方案文件存储到一个新文件中，选择本选项并在语句命令窗口中改变文件名。

单击 Finish 按钮，则还会弹出 SPSS Syntax Editor 窗口，见图 5-13。在该窗口中，对存储的文件名进行修改，单击 Run 菜单，见图 5-14。选择 All 运行，则可将抽样设计方案存储在一个新文件中。



图 5-12 Completion 对话框

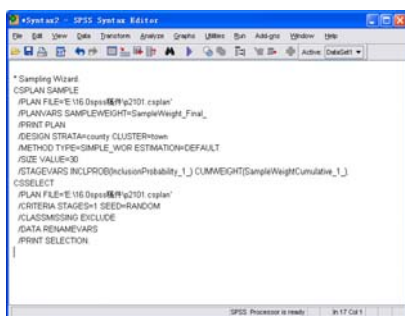


图 5-13 语句命令窗口

(2) 修改

单击 Back 按钮或单击左侧矩形框中上面的某个要点项可对该步骤所作的方案进行及时的更改。

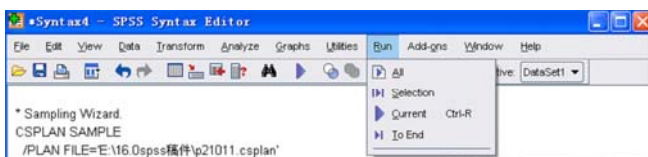


图 5-14 修改文件名并运行

(3) 执行

单击 Finish 按钮，则递交执行。在输出窗中，出现你所设计的方案说明。在原始数据集中出现选中的个体及相关说明。

在本例中，保持系统默认选项，即想存储抽样方案文件，并抽取一个样本。

值得一提的是，如果在前面“6. 定义 Output Variable”中，对输出变量已作过定义，并在本次单击 Finish 按钮前已经运行过中间的抽样设计方案，则在工作的数据文件中，就会生成由前面定义的一部分输出变量，在这种情况下，再次单击 Finish 按钮，系统会自动弹出对冲突的输出变量进行处理的对话框，见图 5-15。

在图 5-15 中,单击 **Rename** 按钮,则对发生冲突的新的输出变量改名后存储到当前工作的数据文件中。

单击 **Delete** 按钮,则先删除当前工作的数据文件中发生冲突的输出变量,再将新的输出变量存储到当前工作的数据文件中。

单击 **Cancel** 按钮,不向当前工作的数据文件中存储新的输出变量。

11. 单击 **Finish** 按钮运行。

12. 运行结果:

在输出窗中,得到输出信息,因信息量较大,故单击输出信息的先后顺序分别列出并单独解释。

(1) 抽样设计方案摘要表,见表 5-3。

表 5-3 抽样设计方案摘要

Summary			
Design Variables	Cluster	1	Stage 1
Sample Information	Selection Method		编号
	Number of Units Sampled		Simple random sampling without replacement
	Variables Created or Modified	Stagewise Inclusion (Selection) Probability	35
		Stagewise Cumulative Sample Weight	Inclusion n (Selection) Probability for Stage 1
		Stagewise Population Size	Cumulative Sampling Weight for Stage 1
		Stagewise Sample Size	Population Size for Stage 1
		Stagewise Sampling Rate	Sample Size for Stage 1
		Stagewise Sample Weight	Sampling Rate for Stage 1
Analysis Information	Estimator Assumption		Sampling Weight for Stage 1
	Inclusion Probability		Equal probability sampling without replacement
			Obtained from variable inclusion n (Selection) Probability for Stage 1

Plan File: E:\2010年书籍\第5章_抽样方法\data5-01.csplan
Weight Variable: Final Sampling Weight

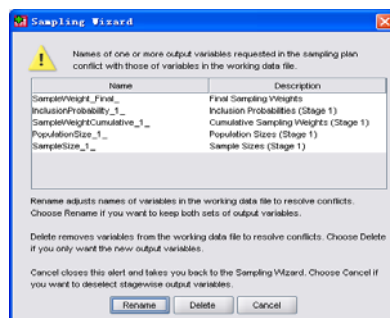


图 5-15 对冲突输出变量的处理

表 5-3 回顾了整个抽样方案，表明抽样设计是一个一阶抽样（只对总体进行一次抽样），列出了在一阶中定义的群变量（编号）及抽样方法（不重复简单随机抽样）、抽样单元的数量(35)、建立或修改的变量：阶段入样(选择)概率(Inclusion Probability_1_)、阶段累计样本权重(Sample Weight Cumulative_1_)、阶段总体容量(Population Size_1_)、阶段样本容量(SampleSize_1_)、阶段抽样比例(Sampling Rate_1_)、阶段样本权重(SampleWeight_1_)，分析信息：估计量假定(等概不重复抽样)、入样概率(从入样概率变量中获取)。你可以用表中的信息来核实这是否就是你头脑中想要的设计的方案。

(2) 第一阶段抽样摘要表，见表 5-4。

表 5-4 第一阶段摘要

Summary for Stage 1

Number of Units Sampled		Proportion of Units Sampled	
Requested	Actual	Requested	Actual
35	35	11.7%	11.7%

PlanFile:E:\2010 年书稿\第 5 章抽样方法\data5-01.csplan

表 5-4 列出了抽样第一阶段的摘要，第一列中显示抽样单元的数量，要求抽取的数量为 35、实际抽取的数量也为 35，第二列列出了层中抽样单元的比例，要求比例和实际比例相等均为 11.7%。

(3) 数据编辑窗口中的抽样结果，见图 5-16。

在数据编辑窗口中，可以得到本次的抽样结果。7 个新增的带下画线的变量被存放到工作的数据文件中。

在 InclusionProbability_1_变量是第一阶段体现的入样概率值。

在 SampleWeightCumulative_1_变量是第一阶段样本权重累积。

PopulationSize_1_变量是总体容量。

SampleSize_1_变量是样本容量。

SamplingRate_1_变量是抽样比例。

SampleWeight_1_变量是样本权重。

在 SampleWeight_Final_变量是最终的样本权重。

在图 5-12 中，含有这 7 个新增变量值的样品被选入样本。而这 7 个新增变量值为系

编号	InclusionProb_1_	SampleWeightCumulative_1_	PopulationSize_1_	SampleSize_1_	SamplingRate_1_	SampleWeight_1_	SampleWeight_Final_			
1	1.00									
2	2.00	0.12	8.57	300	35	0.12	8.57	8.57		
3	3.00									
4	4.00									
5	5.00									
6	6.00									
7	7.00									
8	8.00									
9	9.00									
10	10.00									
11	11.00									
12	12.00									
13	13.00	0.12	8.57	300	35	0.12	8.57	8.57		
14	14.00	0.12	8.57	300	35	0.12	8.57	8.57		
15	15.00									
16	16.00									
17	17.00									
18	18.00									
19	19.00									

图 5-16 在工作数据文件中的新增变量

统缺失值的样品不选入样本。

现在我们可以去调查所选择样本中该社区对应编号的居民每月每户用以食物的消费支出。一旦完成收集工作，就可以用 Complex Samples 分析程序处理样本，使用抽样方案 data5-01.csplan 提供抽样说明。

后缀为 csplan 的文件可以在图 5-2 Welcome to Sampling Wizard 对话框中，选用 Edit a sample design 来查看和编辑。

5.2.3 简单随机抽样的估计

抽样的目的是对总体目标量（研究指标的值）进行估计。

5.2.3.1 简单估值法

1. 均值、总数、方差和均方偏差等的估计

设 y_1, y_2, \dots, y_n 是从总体 $\{Y_1, Y_2, \dots, Y_N\}$ 中抽取的一个容量为 n 的简单随机样本，则样

本均值 $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$ 是总体均值 $\bar{Y} = \frac{\sum_{i=1}^N Y_i}{N}$ 的无偏估计。该估计 \bar{y} 的均方偏差（无偏时即为

方差）为： $V(\bar{y}) = E(\bar{y} - \bar{Y})^2 = \frac{1}{n} \left(1 - \frac{n}{N}\right) S^2$ ，其中总体方差为 $S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2$ 。此

外， $N\bar{y}$ 是总体总数 $Y = \sum_{i=1}^N Y_i$ 的无偏估计，其均方偏差为 $V(N\bar{y}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S^2$ 。样本方

差 $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ 是总体方差 S^2 的无偏估计，而 $v(\bar{y}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) s^2$ 是估计 \bar{y} 的均方偏差 $V(\bar{y})$ 的无偏估计。

2. 实例分析

例 5.5 现以例 5.4 中用“E: \第 5 章 抽样方法\data5-01.csplan”的抽样方案随机抽中的 35 户居民，通过调查获得有关户人数、人均月收入和户月食物支出等方面的调查资料为例（已存放在 data05-02.sav 中），说明在 SPSS 中如何进行均值和均方偏差的估计。

(1) 在数据编辑窗口中，打开 Data05-02.sav。

(2) 按 Analyze→Complex Samples→Descriptives 顺序，展开 Complex Samples Plan for Descriptives 对话框，见图 5-17。要在复式采样分析程序中进行统计分析，首先要有事先建立好的含有分析说明的分析方案文件，或事先建立好的有效的抽样设计方案文件。

① 输入方案文件名

在 Plan 下方的 File 后面的输入框中输入用来进行描述分析的分析方案文件名或抽样设计方案文件名，或单击 Browse 按钮，选取存放分析方案的外部文件名。

本例在 Plan 下方的 File 后面的输入框中输入“E:\第 5 章 抽样方法\data5-01.csplan”。

② 设置联合概率的来源

为了对使用 PPS WOR 方法抽取的群使用 Unequal WOR 估计，你必须指定一个单独的文件或一个打开的包含联合概率的数据集。这个文件或数据集是在抽样期间使用 Sampling Wizard 建立的。

在 Joint Probabilities（联合概率）下面共有三个选项，用来设置设置联合概率的来源。

- Use default file 选项，这是系统默认选项，选择本选项，是指使用在 File 中输入的文件。

- An open dataset 选项，选择此项，可以从其下框中选择一个打开的包含联合概率的数据集。

- Custom file 选项，选择此项，可在其下面的 File 后框中输入或单击 Browse 按钮，自定义一个用来进行描述分析的分析方案文件或抽样设计方案文件。

在本例中使用系统默认选项，即在建立抽样设计方案文件时用到并存放抽样数据的“E:\第 5 章 抽样方法\data05-01.sav”数据文件。

(3) 单击 Continue 按钮，进入如图 5-18 所示的描述统计对话框。



图 5-17 Descriptives 复合采样方案对话框

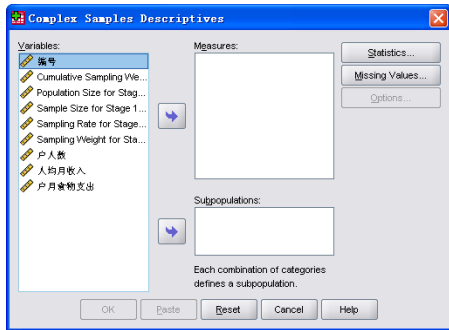


图 5-18 描述统计对话框

① 定义测量变量

在左侧变量名列表中至少选择一个测量变量，将其移入到 Measures 下框中。测量变量应是尺度变量。

本例在左边变量名列表中，选择户月食物支出并将其移入到 Measure 下框中，作为测量变量。

② 定义子总体变量

同样地，在左侧变量名列表中，选择变量并将其移入到 subpopulations 下框中，来定

义子总体。这样可以为每个子总体单独计算统计量。子总体变量可以是字符串或数字型且应是分类的。

本例对此不做任何定义。

(4) 单击 statistics 按钮, 展开 Statistics 对话框, 见图 5-19。在本对话框中可以选择统计量。

① Summaries 栏, 共有 4 个选择项, 系统默认选项为 Mean。

- Mean 选项, 选择本项要求做均数估计。

- t-test 选项, 位于 Mean 选项下的 t-test 选项只有当选择 Mean 选项后, 处于激活状态。选择本选项, 可以做均数估计与指定值的 t 检验。在 Test Value 后框中输入检验的一个指定值。

- Sum 选项, 选择本项要求对测量变量做总数估计。

- t-test 选项, 位于 Sum 选项下的 t-test 选项只有当选择 Sum 选项后, 处于激活状态。选择本选项, 可以做总和估计与指定值的 t 检验。在 Test Value 后框中输入检验的一个指定值。

② Statistics 栏, 它产生同均数和总数有关的统计量。

- Standard error。选择此项, 产生估计的标准误差。

- Confidence interval。选择此项, 可在其下面的 level (%) 框中输入置信区间的水平值, 默认值为 95, 从而可产生使用指定水平的估计的置信区间。

- Coefficient of variation。选择此项, 可求变异系数。

- Unweighted count。选择此项, 可以产生计算估计的单元数, 即样本量。

- Population size。选择此项, 可以产生总体容量。

- Design effect。选择此项, 可以产生一个估计的方差与基于假定的一个简单随机样本获得的方差的比值, 称为设计效应。它是指定的一个复合设计效应的测度, 值比 1 越小表明效应越好。

- Square root of design effect。选择此项, 可以产生一个设计效果的平方根值, 它也是指定的一个复合设计效果的测度, 同样该值比 1 越小表明效果越好。

从中你可选择一项、多项甚至可全部选定。

本例在 Statistics 选项中, 除默认选项以外, 再选择 Confidence interval、Unweighted count、Population size、Design effect 选项。其他保持系统默认选项。

单击 Continue 按钮, 返回 Complex Samples Descriptives 对话框。

(5) 单击 Missing Values 按钮, 展开描述统计 Missing Values 对话框, 见图 5-20。选择对缺失值的处理方法。

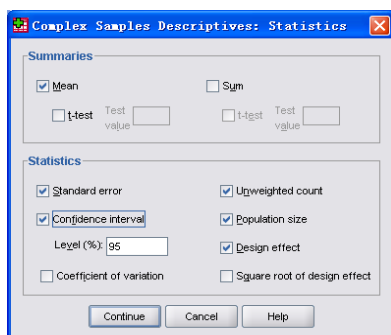


图 5-19 描述统计对话框

- **Statistics for Measure Variables** 栏，本组选择项决定哪些样品使用到分析中。
- **Use all available data** 选项，缺失值是由各变量逐一确定的，因此，用来计算统计量的样品在所有的测量变量中是可变的。

- **Ensure consistent case base** 选项，缺失值是根据所有变量确定的，因此用来计算各个统计量的样品数是一致的。

② **Categorical Design Variables** 栏，确定用户缺失值是否有效或无效。它提供两个选项需要选择：

- **user-missing values are valid** 选项，选择本项，用户的缺失值是有效的。它是系统默认选择项。

- **user-missing values are invalid** 选项，选择本项，用户的缺失值是无效的。

单击 **Continue** 按钮，返回 **Complex Samples Descriptives** 对话框。

由于本例只有一个变量，故本例对缺失值的处理采用系统默认选项。

(6) 单击 **Options** 按钮，展开 **Options** 对话框，参见图 5-21。选择 **Options** 任选项。

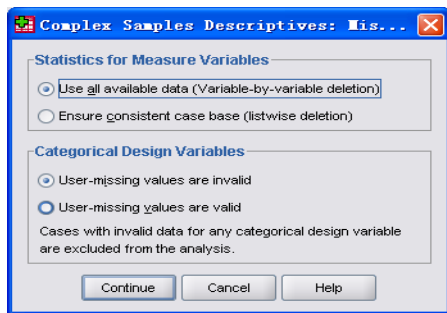


图 5-20 描述统计 Missing Values 对话框

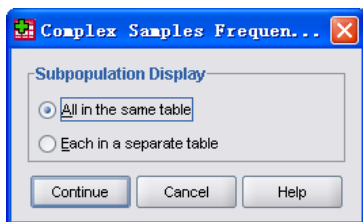


图 5-21 Options 对话框

Subpopulation Display 栏中有两个选项：

① **All in the same table** 选项，它是系统的默认选项，选项此项意为所有的子总体在同一张表中显示。

② **Each in a separate table** 选项，选择此项，则每个子总体在一个单独的表中显示。

由于本例只有一个变量，故本例对缺失值的处理采用系统默认选项。即所有的子总体在同一张表中显示。

单击 **Continue** 按钮，返回 **Complex Samples Descriptives** 对话框。

(7) 单击 **Paste** 按钮，可在 **Syntax** 窗中生成本程序的语句。参见图 5-13。

(8) 单击 **OK** 按钮，执行运算，在输出窗中得到输出结果。

(9) 结果解释：

在输出窗中，产生一张表，见表 5-5。

表 5-5 对户月食物支出变量计算的各个选定的统计量

Univariate statistics							
	Estimate	Standard Error	95%Confidence Interval		Design Effect	Population Size	Unweighted Count
			Lower	Upper			
Mean 户月食物支出	8.9571E2	34.59441	825.4100	966.0186	1.000	300.000	35

表 5-5 对户月食物支出变量计算各个选定的统计量。第一列是户月食物支出的均值估计。第二列是其对应估计的标准误差。第三列是估计的 95%的置信区间的下限和上限。第四列列出的是方案的设计效应 (Deff)，它是由任一种设计的抽样方案确定的总体均值（或总数）的估计量的均方偏差与简单随机抽样的简单估值法确定的估计量的均方偏差之比。由于本例中两个方法都是一样的，故其值为 1。第五列列出的是总体容量，本例为 300。第六列列出的是抽取的样本量为 35。

结论：用 35 户的一个简单随机样本的调查结果，可以估计总体的户月食物支出的均值为 895.71 元，标准误差为 34.59 元，总体户月食物支出的均值（元）的 95%的置信区间为 (825.42, 966.02)。

例 5.6 某商店有 100 个货架，为节省盘点时间，现用简单随机抽样的方法，抽取 45 个货架上的货物进行清点，并依此记录抽中的 45 个货架上的货物价值，请从得到的样本来估计总的货物价值及其估计值的误差和 95%的置信区间，并检验可否认为该商店的货物总额约为 60000 元。

在 SPSS 中，基本的操作步骤如下：

- ① 在 SPSS 中数据编辑窗中，先建立 100 个货架的抽样框数据文件，见 data05-06.sav。
- ② 单击 Analyze→Complex Samples→Select a Sample 顺序，打开 Welcome to Sampling Wizard 对话框，见图 5-2。
- ③ 在 Sampling Wizard 对话框中，选择 Design a sample 选项，将插入点定位在其后 File 的文本编辑框中，输入“E:\第 5 章 抽样方法\data5-02.csplan”。定义抽样方案文件名。
- ④ 单击 Next 按钮，进入 Stage 1: Design Variables 对话框。
- ⑤ 在中间的变量名列表中选择货架编号，单击右侧第二个右移箭头，将货架编号移入到 Cluster 下框中，定义货架编号为第一阶群变量。
- ⑥ 单击 Next 按钮，进入 Stage 1: Sampling Method 对话框。在本对话框中，不作任何选择，采用系统默认的 Simple Random Sampling（简单随机抽样法）。
- ⑦ 单击 Next 按钮，打开 Stage 1: Sample Size 对话框。在本对话框中，定义样本容量。在 Units 的下拉式列表中选择 Counts 来定义抽样单位为总数。在 Values 下框中输入

45. 表示在本阶段中选择 45 个样本单元。

⑧ 单击 Next 按钮, 进入 Stage 1: Output Variable 对话框, 见图 5-8。在输出变量对话框中不做任何选择。

⑨ 单击 Next 按钮, 进入 Stage 1: Plan Summary 对话框, 见图 5-9。查看一下抽样设计。保持系统默认选择, 不再增加其他阶的抽样。

⑩ 单击 Next 按钮, 进入到 Stage 2: Selection Options 对话框, 见图 5-10。保持系统默认选择。

⑪ 单击 Next 按钮, 进入到 Stage 2: Output Files 对话框, 见图 5-11。选择 External file 选项, 输入 “E:\第 5 章 抽样方法\data05-07.sav”, 即将样本数据存在一个外部的数据文件中。

⑫ 单击 Next 按钮, 进入到 Stage 2: Completion 对话框, 见图 5-12。保持系统默认选项, 表示你想存储抽样方案文件, 并抽取一个样本。

⑬ 单击 Finish 按钮运行。

则在输出窗中出现抽样设计方案摘要表和第一阶段抽样摘要表, 参见例 5.4 中对表 5-3 和表 5-4 的解释。

抽样数据存放在 data05-07.sav 中, 在数据编辑窗口中打开它, 可见到本次的抽样结果, 包括抽中的货架编号和存放第一阶段包含概率值的 InclusionProbability_1_、第一阶段累积样本权重 SampleWeightCumulative_1_及存放最终样本权重的 SampleWeight_Final_。

⑭ 记录这 45 个货架的货物的价值, 并输入到 data05-07.sav 中新增变量价值一列的对应行中。另存为 data05-07a.sav。

注: 实际工作中可直接存在 data05-07.sav 中。

⑮ 按 Analyze→Complex Samples→Descriptives 顺序, 展开 Complex Samples Plan for Descriptives 对话框, 见图 5-17。

⑯ 在 Plan 下方的 File 后面的输入框中输入 “E:\第 5 章 抽样方法\data5-02.csplan”。由于采用的是简单随机抽样, 因此, 其他不做任何选择。

⑰ 单击 Continue 按钮, 进入如图 5-18 所示的描述统计对话框。

在左边变量名列表中, 选择 价值 并将其移入到 Measure 下框中, 作为测量变量。

⑱ 单击 Statistics 按钮, 展开如图 5-19 所示的描述统计对话框。

在 Summaries 中, 关闭系统默认选项 Mean, 选择 Sum 选项, 选中其下的 t-test, 并在其后框中输入 60000, 要求对所估计的总数与 60000 做标准对照的 t 检验。

在 Statistics 选项中, 除默认选项外, 另外选择 Confidence interval、Coefficient of variation、Unweighted count、Population size、Design effect 选项。

⑲ 单击 Continue 按钮, 返回 Complex Samples Descriptives 对话框。

⑳ 单击 OK 按钮, 执行运算, 在输出窗中得到输出结果。

㉑ 结果解释:

在输出窗中, 产生一张表, 见表 5-6。

表 5-6 对价值变量计算的各个选定的统计量

Univariate statistics								
	Estimate	Standard Error	95%Confidence Interval		Hypothesis Test			
			Lower	Upper	Test Value	t	df	Sig
Sum 价值	5.86E4	1222.73010	56129.08	61057.58	60000	-1.150	44	256

表 5-6 对价值变量计算各个选定的统计量。第一列是求总和的变量名(价值), 第二列是价值总和的估计, 第三列是其对应估计的标准误差, 第四列是估计的 95%的置信区间的下限和上限。第五列是假设检验(依次为检验值、t 值、自由度、无差异的原假设成立的概率值), 第六列列出的是变异系数的值, 第七列列出的是方案的设计效应, 由于本例中两个方法都是(简单随机抽样)一样的, 故其值为 1。第八列列出的是总体容量, 本例为 100。第九列列出的是抽取的样本量, 本例为 45。

结论: 用 45 个货架的一个简单随机样本的调查结果, 可以估计 100 个货架总体的总价值为 58600 元, 标准误差为 1222.73 元, 总价值的 95%的置信区间为: (56129.08, 61057.58), 单位: 元。该商店的货物总额等于 60000 元原假设下, 出现目前统计量的值或者更极端值的概率为 0.256, 大于 0.05, 故没有足够证据拒绝该商店的货物总额等于 60000 元的原假设。

5.2.3.2 比估计

1. 比估计和均方偏差估计

在抽样调查中在两种情况下要用到比估计, 一种情况是所估计的目标值是两个指标总数之比 ($R = \frac{Y}{X}$); 另一个情况是所估计的目标值是某指标 Y 的总数, 但有另一与 Y 关系密切的指标 X 可作为辅助变量, 利用辅助变量的信息可以改进估计的精度。

设有目标量 Y (指标 1) 和辅助量 X (指标 2), 要估计 $R = \frac{\bar{Y}}{\bar{X}} = \frac{Y}{X}$, Y 、 X 为两个指标的总数, 可用 $r = \bar{y}/\bar{x} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$ 估计 R 。以 r 估计 R 是近似无偏的, 其均方偏差近似为: $E(r - R)^2 = \frac{1-f}{n} \frac{1}{\bar{X}^2} \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2$ 。其中 $f = \frac{n}{N}$, n 为抽取的样本量, N 是总体容量。当 \bar{X} 已知时, 则可用 $\bar{y}_R = r\bar{X}$ 估计 \bar{Y} , 其均方偏差近似为

$$V(\bar{y}_R) = \bar{X}^2 E(r - R)^2 = \frac{1-f}{n} \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2$$

其估计量可采用

$$v(\bar{y}_R) = \frac{1-f}{n} \frac{1}{n-1} \sum_{i=1}^n (y_i - rx_i)^2$$

2. 实例分析

例 5.7 从某地区 15786 位老人中（抽样框在 data05-08.sav 中），用简单随机抽样抽取一个容量为 525 的随机样本，抽样设计方案存放在 data5-03. csplan 文件中，现对 525 名抽中的老人进行性别调查，调查结果存放在 data05-09.sav 中，使对其求比估计。

本例是个特例，有两种解法。

解法一的步骤如下：

- ① 在数据编辑窗口中，打开 Data05-09.sav。
- ② 按 Analyze→Complex Samples→Descriptives 顺序，展开 Complex Samples Plan for Descriptives 对话框，见图 5-17。
- ③ 在 Plan 下方的 File 后面的输入框中输入“E:\第 5 章 抽样方法\data5-03.csplan”。
- ④ 单击 Continue 按钮，进入如图 5-18 所示的描述统计对话框。
在左边变量名列表中，选择性别并将其移入到 Measure 下框中，作为测量变量。
- ⑤ 单击 Statistics 按钮，展开如图 5-19 所示的描述统计对话框。
在 Statistics 选项中，除默认选项以外，再选择 Confidence interval、Unweighted count、Population size、Design effect 选项。其他保持系统默认选项。
- ⑥ 单击 Continue 按钮，返回 Complex Samples Descriptives 对话框。
- ⑦ 单击 OK 按钮，执行运算，在输出窗中得到输出结果。
- ⑧ 结果解释：
在输出窗中，产生一张表，表 5-7。

表 5-7 对性别变量计算的各个选定的统计量

Univariate Statistics							
	Estimate	Standard Error	95%Confidence Interval		Design Effect	Population Size	Unweighted Count
			Lower	Upper			
Mean 性别	0.4610	0.02141	0.4189	0.5030	1.000	15786.000	525

表 5-7 中各列的含义同例 5.6 中的结果解释。由于性别男设置为 1 而性别女设置为 0，因而总体均值的估计值 0.4610，实际上就是男性在总体中的比率的估计值。

解法二的步骤如下:

① 在数据编辑窗口中, 打开 Data05-09.sav, 并增加一列新的常量, 变量名为参与调查, 对 525 人的赋值均为 1。

② 按 Analyze→Complex Samples→Ratios 顺序, 展开 Complex Samples Plan for Ratios 对话框, 同图 5-17 十分类似。在 Plan 下方的 File 后面的输入框中输入“E:\第5章 抽样方法\data5-03.csplan”。其他不做任何选择。

③ 单击 Continue 按钮, 进入如图 5-22 所示的 Complex Samples Ratios 对话框。

- 设置分子变量

在左侧变量名列表中, 选择有正值的尺度变量, 并将其移入到 Numerators 的下框中。定义行变量。

- 设置分母变量

在左侧变量名列表中, 选择有正值的尺度变量, 并将其移入到 Denominators 的下框中。定义列变量。

- 定义分组变量

根据需要, 在左侧变量名列表中, 选择字符串或数字型的分类变量, 并将其移入到 Subpopulations 下框中。

本例, 在左边变量名列表中, 选择性别并将其移入到 Numerators 下框中, 作为分子变量; 选择参与调查并将其移入到 Denominator:下框中, 作为分母变量。

④ 单击 Statistics 按钮, 展开如图 5-23 所示的比率统计对话框。

本对话框中各统计量的说明, 请参阅例 5.5 中的相关内容。

本例, 在 Statistics 选项中, 除默认选项以外, 再选择 Confidence interval、Unweighted count、Population size、Design effect 选项。其他保持系统默认选项。

⑤ 单击 Continue 按钮, 返回 Complex Samples Ratios 对话框。

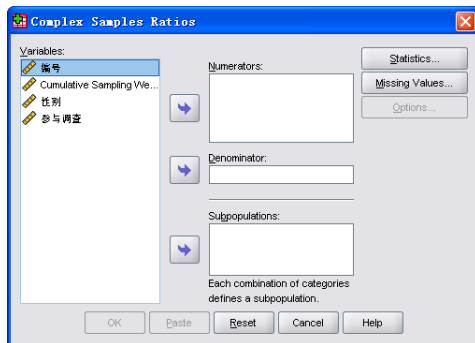


图 5-22 Complex Samples Ratios 对话框

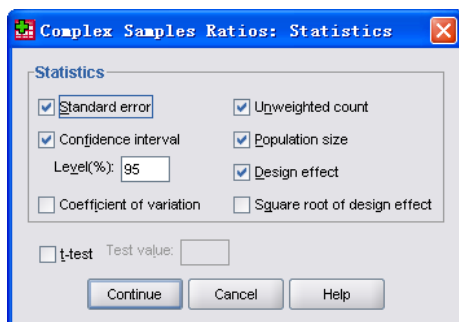


图 5-23 比率统计对话框

⑥ 单击 OK 按钮，执行运算，在输出窗中得到输出结果。

⑦ 结果解释：

在输出窗中，产生一张表，表 5-8。

表 5-8 对性别变量计算的各个选定的统计量

Ratios 1								
Num erator	Denomi nator	Ratio Estimate	Standard Error	95%Confidence Interval		Design Effect	Population Size	Unweight ed Count
				Lower	Upper			
性别	参与调查	0.461	0.021	0.419	0.503	1.000	15786.000	525

表 5-8 对性别变量计算各个选定的统计量。第一列列出的分子和分母的变量名，第二列列出的是性别比的估计。第三列列出的是其对应估计的标准误差。第四列列出的是估计的 95%的置信区间的下限和上限。第五列列出的是方案的设计效应，由于本例中两个方法都是（简单随机抽样）一样的，故其值为 1。第六列列出的是总体容量，本例为 15786。第七列列出的是抽取的样本量，本例为 525。

结论：对 525 名老人的一个简单随机样本的性别调查结果，可以估计男性比率为 0.461，标准误差为 0.021，总体中男性比率的 95%的置信区间为 (0.419, 0.503)。

例 5.8 仍以例 5.4 中用“E: \第 5 章 抽样方法\data5-01.csplan”的抽样方案随机抽中的 35 户居民，通过调查获得有关户人数、人均月收入 and 户月食物支出等方面的调查资料为例（已存放在 data05-02.sav 中），说明在 SPSS 中如何用比估计法进行户月食物支出均值和均方偏差的估计。

完整的步骤如下：

① 在数据编辑窗口中，打开 data05-02.sav。

② 按 Analyze→Complex Samples→Ratios 顺序，展开 Complex Samples Plan for Ratios 对话框，同图 5-17 十分类似。

③ 在 Plan 下方的 File 后面的输入框中输入“E: \第 5 章 抽样方法\data5-01.csplan”。

④ 单击 Continue 按钮，进入如图 5-12 所示的 Complex Samples Ratios 对话框。

在左边变量名列表中，选择户月食物支出并将其移入到 Numerators:下框中，作为分子变量；选择户人数并将其移入到 Denominator 下框中，作为分母变量。

⑤ 单击 Statistics 按钮，展开如图 5-13 所示的比率统计对话框。

在 Statistics 选项中，除默认选项以外，再选择 Confidence interval、Unweighted count、Population size、Design effect 选项。其他保持系统默认选项。

⑥ 单击 Continue 按钮，返回 Complex Samples Ratios 对话框。

⑦ 单击 OK 按钮，执行运算，在输出窗中得到输出结果。

⑧ 结果解释:

在输出窗中,产生一张表,表 5-9。

表 5-9 对户月食物支出变量与户人数变量比计算的各个选定的统计量

Ratios 1								
Num erator	Denomi nator	Ratio Estimate	Standard Error	95%Confidence Interval		Design Effect	Population Size	Unweighted Count
				Lower	Upper			
用户食物支出	户人数	261.250	6.213	248.625	273.875	1.000	300.000	35

表 5-9 对户月食物支出变量与户人数变量比计算各个选定的统计量。第一列列出的分子和分母的变量名,第二列列出的是户月食物支出变量与户人数变量比的估计。第三列列出的是其对应估计的标准误差。第四列列出的是估计的 95%的置信区间的下限和上限。第五列列出的是方案的设计效应,由于本例中两个方法都是(简单随机抽样)一样的,故其值为 1。第六列列出的是户数的总体容量,本例为 300。第七列列出的是抽取的样本量,本例为 35。

因此,用比估值法来估计平均每月用以食物的支出,可得估计值为

$$\bar{y}_R = r\bar{X} = 261.25 \times \frac{1100}{300} = 957.92 \text{ (元)}$$

这一估计的均方偏差的估计值为

$$v(\bar{y}_R) = \frac{1-f}{n} \frac{1}{n-1} \sum_{i=1}^n (y_i - rx_i)^2 = 6.213^2 \times \bar{x}^2 = 6.213^2 \times \left(\frac{120}{35}\right)^2 = 453.763$$

所以,标准误差为: $\sqrt{v(\bar{y}_R)} = \sqrt{453.763} = 21.30$ 。

以上结果同简单估计法例 5.5 的结果相比,估计的平均每月用以食物的支出略大(895.71 元)、而标准误差要小得多(34.59441)。

简单随机抽样推算得到样本含量还可用来估计其他抽样方法所需的样本含量。其他抽样方法所需的样本含量可通过下述方法得出:

- (1) 在给定一定精度下,先确定简单随机抽样简单估值法时所需的样本含量 n_0 ;
- (2) 估算抽样实施方案的设计效应值 $Deff$;
- (3) 实施方案所需的样本含量: $n=n_0 \cdot (Deff)$ 。

5.3 系统随机抽样

5.3.1 系统随机抽样概述

系统随机抽样(Systematic Random Sampling) 又称机械抽样或等距抽样。它是把总体中所有单元(个体)按某种顺序排列编号,然后以固定的间隔抽取样品的一种抽样方法。

习惯地用 n 表示所要抽取的样本容量,样本容量的计算方法,可依 5.2.4。用 N 表示总体容量。则用 $k=\text{int}(N/n)$,表示求总体容量除以样本容量的整数。那么,系统随机抽样的具体做法为(1)随机地从抽样框前 k 个名录中选择一个被试对象,以及(2)在此之后选择每隔第 k 个列出的被试对象。数量 k 被称为跳过的数量,即间隔。

例如,你想要从一个校园学生名单目录里列出的 10008 名学生的总体中得到 100 名学生的一个系统随机样本。则此时的样本容量 $n=100$ 和总体容量 $N=10008$,因此, $k=\text{int}(10008/100)=100$ 。由于总体容量是样本容量的 100 倍,所以你需要从每 100 名学生中选择一名。首先从目录的前 100 名学生中,使用随机数,随机地选择一名学生。然后在随机选择的那名学生之后,选择每隔第 100 名的学生。这产生容量 100 的一个样本。

从一个抽样框中抽样,选择一个系统随机样本要比选择一个简单随机样本更简单,因为它只使用一个随机数。这种方法典型地为总体提供了一个良好的代表性样本,因为对于字顺表如名单目录,大多数变量的值在名单列表上是随机地波动的。

5.3.2 系统随机抽样在 SPSS 中的实现

例 5.9 某单位共有 550 名职工,现想用系统随机抽样方法从中随机抽取 30 名职工的一个样本,调查其工资外收入情况,以财务职工编号为作为抽样框,抽取一个系统随机样本,并对其进行调查,依此估计 550 名职工总体的工资外平均收入情况。

在 SPSS 中的操作步骤如下:

① 在 SPSS 中数据编辑窗中,打开 data05-10.sav。

② 单击 Analyze→Complex Samples→Select a Sample 顺序,打开如图 5-2 所示的 Welcome to Sampling Wizard 对话框。在 Sampling Wizard 对话框中,选择 Design a sample 选项,将插入点定位在其后 File 的文本编辑框中,输入“E:\第 5 章抽样方法\data5-04.csplan”。定义抽样方案文件名。

③ 单击 Next 按钮,进入 Stage 1: Design Variables 对话框。在中间的变量名列表中选择职工编号,单击右侧第二个右移箭头,将职工编号移入到 Cluster 下框中,定义职工编号为第一阶群变量。

④ 单击 Next 按钮,进入 Stage 1: Sampling Method 对话框。在 Mothed 的 Type 下拉

式选项中选择 Simple Systematic (简单系统抽样法)。其他采用系统默认的 without replacement(WOR)选项,即在本阶段中,利用默认的不重复(WOR)简单系统抽样法,抽取的职工编号作为主要的抽样单元。

⑤ 单击 Next 按钮,打开 Stage 1: Sample Size 对话框。在本对话框中,定义样本容量。它在 Units 的下拉式列表中通过选择 Counts 或 Proportions 来定义抽样单位为总数还是比例。本例选择 Counts,因此,在 Values 下框中输入 30。表示在本阶段中选择 30 个样本单元。

⑥ 单击 Next 按钮,进入 Stage 1: Output Variable 对话框,见图 5-8。不做任何选择。

⑦ 单击 Next 按钮,进入 Stage 1: Plan Summary 对话框,见图 5-9。查看一下抽样设计。保持系统默认选择,不再增加其他阶的抽样。

⑧ 单击 Next 按钮,进入到 Stage 2: Selection Options 对话框,见图 5-10。采用系统默认选择。

⑨ 单击 Next 按钮,进入到 Stage 2: Output Files 对话框,见图 5-11。选择 External file 选项,输入“E:\第 5 章 抽样方法\data05-11.sav”,即将样本数据存在一个外部的数据文件中。

⑩ 单击 Next 按钮,进入到 Stage 2: Completion 对话框,见图 5-12。保持系统默认选项,去存储抽样方案文件,并抽取一个样本。

⑪ 单击 Finish 按钮运行。

则在输出窗中出现抽样设计方案摘要表和第一阶段抽样摘要表,参见例 5.4 中对表 5-3 和表 5-4 的解释。

5.3.3 系统随机抽样的估计

1. 基本估计方法

如果总体单元的排列顺序是完全随机的,则系统抽样可当作简单随机抽样处理,但事实上,一个系统随机样本不是一个简单随机样本,因为所有的容量 n 的样本不是等可能的。比如,它与一个简单随机样本不同的是在名单列表中列出的彼此相邻的两个被试对象是不能同时出现在样本中。

系统抽样的模型可表述如下:选择一个正整数 K ,将总体中 N 个个体单元依次排列为

1,	2,	...	K ,
$K+1$,	$K+2$,	...	$2K$,
$2K+1$,	$2K+2$,	...	$3K$,
.....			
直到 N 为止,			

然后对号码 1, 2, ..., K 作随机抽样。当我们把同一列的个体看作一个群, 则 N 个个体分成 K 个群, 且 $N_1 = N_2 = \dots = N_K = N_0 = \frac{N}{K}$, 若 i 入样, 则相当于所在 i 列的整群入样, 所以, 系统抽样法可视为整群抽样中抽取第一阶样本单元数为 1 的抽样。

所以, 可用

$$\hat{Y}_{\text{sys}} = K \sum_{j=1}^{N_0} Y_{\theta j}$$

估计总体总数 Y , 其中 θ 为从 1, 2, ..., K 中随机抽取的入样号码。 $Y_{\theta j}$ 是第 i 列, 第 j 行的个体单元。其估计的均方偏差为

$$V(\hat{Y}_{\text{sys}}) = K^2 \left(1 - \frac{1}{K}\right) \frac{1}{K-1} \sum_{i=1}^K \left(\sum_{j=1}^{N_0} Y_{ij} - \frac{Y}{K} \right)^2 = \frac{N^2}{K} \sum_{i=1}^K (\bar{Y}_i - \bar{Y})^2$$

对均方偏差 $V(\hat{Y}_{\text{sys}})$ 的估计的方法有很多, 一般都是有偏估计, 其中之一为

$$v(\hat{Y}_{\text{sys}}) = N^2 \sum_{j=1}^{N_0/2} \left(\frac{2K}{N} \right)^2 \frac{1}{2} \left(1 - \frac{2}{2K} \right) \frac{(Y_{\theta(2j)} - Y_{\theta(2j-1)})^2}{2} = K(K-1) \sum_{j=1}^{N_0/2} (Y_{\theta(2j)} - Y_{\theta(2j-1)})^2$$

它是将二

行的个体单元作为一个层, 每层有两个样本单元, 两个样本单元构造该层的一个方差估计, 再按分层抽样汇总出一个均方偏差的估计。

2. 实例分析

例 5.10 接例 5.9 我们调查了所选择样本中该校对应职工编号的职工的工资外收入, 并将资料录入并存放在 data05-11.sav 中。现使用抽样方案 data5-04.csplan 提供抽样说明, 用 Complex Samples 分析程序估计 550 名职工总体的工资外平均收入情况。

① 在数据编辑窗口中, 打开 data05-11.sav。

② 按 Analyze→Complex Samples→Descriptives 顺序, 展开 Complex Samples Plan for Descriptives 对话框, 见图 5-17。在 Plan 下方的 File 后面的输入框中输入“E:\第 5 章 抽样方法\data5-04.csplan”。

③ 单击 Continue 按钮, 进入如图 5-18 所示的描述统计对话框。

在左边变量名列表中, 选择工资外收入并将其移入到 Measure 下框中, 作为测量变量。

④ 单击 Statistics 按钮, 展开如图 5-19 所示的描述统计选项卡。

在 Statistics 选项中, 除默认选项以外, 再选择 Confidence interval、Unweighted count、

Population size、Design effect 选项。其他保持系统默认选项。

⑤ 单击 Continue 按钮，返回 Complex Samples Descriptives 对话框。

⑥ 单击 OK 按钮，执行运算，在输出窗中得到输出结果。

⑦ 结果解释：

在输出窗中，产生一张表，见表 5-10。

表 5-10 对工资外收入变量计算的各个选定的统计量

Univariate statistics							
	Estimate	Standard Error	95%Confidence Interval		Design Effect	Population Size	Unweighted Count
			Lower	Upper			
Mean 工资外收入	1.9887E3	113.96585	1755.5803	2221.7530	1.058	550.000	30

表 5-10 中各列的含义同例 5.5 中的结果解释。由此可见用系统随机抽样得到的 30 名职工样本估计的 550 职工总体的平均工资外收入为 1988.7 元，标准误为 113.96585 元，550 名职工总体的平均工资外收入的 95%的置信区间为 (1755.58, 2221.75)。设计效应为 1.058。

5.4 PPS 抽样

5.4.1 PPS 抽样概述

所谓 PPS 抽样是指抽取概率正比于规模测度的一种不等概率抽样方法。常用于整群抽样中。设总体 $\{U_1, U_2, \dots, U_N\}$ 每个单元有一个指标 Y_i ，还有一个规模测度变量 $X_i > 0$ ， $i = 1, 2, \dots, N$ 。在抽取样本单元时，各单元抽取的概率正比于其规模测度。在有放回 PPS 抽样中，每次抽取，第 i 个单元 U_i 被抽中的概率 p_i 正比于 X_i ，即

$$p_i = X_i / \sum_{i=1}^N X_i$$

一次抽取后，放回被抽中的单元再作下次抽取。

5.4.2 PPS 抽样在 SPSS 中的实现

例 5.11 某地区欲调查奶牛的存栏头数，该地区共有 18 个县，各县的奶牛场数目见表 5-11。考虑到各县奶牛场数目的不同，现想用 PPS 抽样法抽取 6 个县，在 SPSS 中应如何做？结果怎样？

表 5-11 各县的奶牛场数目

县序号	农场数	县序号	农场数	县序号	农场数
1	29	7	35	13	61
2	33	8	30	14	64
3	28	9	41	15	68
4	58	10	48	16	68
5	49	11	53	17	39
6	40	12	55	18	47

在 SPSS 中，用 PPS 进行随机抽样的步骤如下：

① 在 SPSS 中数据编辑窗中，建立变量县序号和农场数两个变量，并将表 5-11 中的数据依次录入到这两个变量中，并将结果存放在 data05-12.sav 中。

② 单击 Analyze→Complex Samples→Select a Sample 顺序，打开如图 5-2 所示的 Welcome to Sampling Wizard 对话框。在 Sampling Wizard 对话框中，选择 Design a sample 选项，将插入点定位在其后 File 的文本编辑框中，输入“E:\第 5 章抽样方法\data5-05.csplan”。定义抽样方案文件名。

③ 单击 Next 按钮，进入 Stage 1: Design Variables 对话框。在中间的变量名列表中选择县序号，单击右侧第二个右移箭头，将县序号移入到 Cluster 下框中，定义县序号为第一阶群变量。

④ 单击 Next 按钮，进入 Stage 1: Sampling Method 对话框。在 Method 的 Type 下拉式选项中选择 PPS。其他采用系统默认的 without replacement(WOR)选项，即在本阶段中，利用默认的不重复(WOR)PPS 抽样法，抽取的县序号作为主要的抽样单元。在 Measure of Size(MOS)下，选择 Read from variable 选项，并将变量名源列表中农场数变量用右移箭头移到 Read from variable 的下框中，定义该变量为规模测度变量。

⑤ 单击 Next 按钮，打开 Stage 1: Sample Size 对话框。在本对话框中，定义样本容量。它在 Units 的下拉式列表中通过选择 Counts 来定义抽样单位为总数。在 Values 下框中输入 6。表示在本阶段中选择 6 个整群。

⑥ 单击 Next 按钮，进入 Stage 1: Output Variable 对话框，见图 5-8。不做任何选择。

⑦ 单击 Next 按钮，进入 Stage 1: Plan Summary 对话框，见图 5-9。查看一下抽样设计。保持系统默认选择，不再增加其他阶的抽样。

⑧ 单击 Next 按钮，进入到 Stage 2: Selection Options 对话框，见图 5-10。采用系统默认选择。

⑨ 单击 Next 按钮，进入到 Stage 2: Output Files 对话框，见图 5-11。选择 External file 选项，输入“E:\第 5 章 抽样方法\data05-13.sav”，即将样本数据存在一个外部的数据文

件中。由于选择了 PPS 抽样法, 因此, 在 Where do you want to save joint probabilities 选择项中, 必须指定联合概率的存储位置。在 file 后框中, 输入 “E:\第 5 章 抽样方法\data05-14.sav”, 即将联合概率存在一个外部的数据文件中。

⑩ 单击 Next 按钮, 进入到 Stage 2: Completion 对话框, 见图 5-12。保持系统默认选项, 去存储抽样方案文件, 并抽取一个样本。

⑪ 单击 Finish 按钮运行。

则在输出窗中出现抽样设计方案摘要表和第一阶段抽样摘要表, 参见例 5.4 中对表 5-3 和表 5-4 的解释。

抽样结果: 抽中的 6 个县的抽样数据资料被存放在 “E:\第 5 章 抽样方法\data05-13.sav” 中。

例 5.12 现将例 5.11 中的存放在 data05-12.sav 数据结构做一些改变, 只有一个县序号变量, 则 1 号县中 29 个农场数可用 29 个 1 来表示 (这样设定不是唯一的, 可以有别的做法), 同样, 2 号县中 33 个农场数可用 33 个 2 来表示, 依次类推, 建立的数据文件存放在 data05-15.sav 中, 在这种情况下, 在 SPSS 中应如何用 PPS 抽样法抽取 6 个县? 结果是如何显示的?

在 SPSS 中, 用 PPS 进行随机抽样的步骤如下:

① 在 SPSS 中数据编辑窗中, 打开 data05-15.sav 数据文件。

② 单击 Analyze→Complex Samples→Select a Sample 顺序, 打开如图 5-2 所示的 Welcome to Sampling Wizard 对话框。在 Sampling Wizard 对话框中, 选择 Design a sample 选项, 将插入点定位在其后 File 的文本编辑框中, 输入 “E:\第 5 章 抽样方法\data5-06.csplan”。定义抽样方案文件名。

③ 单击 Next 按钮, 进入 Stage 1: Design Variables 对话框。在中间的变量名列表中选择县序号, 单击右侧第二个右移箭头, 将县序号移入到 Cluster 下框中, 定义县序号为第一阶群变量。

④ 单击 Next 按钮, 进入 Stage 1: Sampling Method 对话框。在 Mothed 的 Type 下拉式选项中选择 PPS。其他采用系统默认的 without replacement(WOR)选项, 即在本阶段中, 利用默认的不重复 (WOR) PPS 抽样法, 抽取的县序号作为主要的抽样单元。在 Measure of Size(MOS)下, 选择 Count data records 选项, 在 Minimum 后框中输入容量的最小值 29, 在 Maximum 后框中输入容量的最大值 68。以此来任意地设置 MOS 的下限和上限。

⑤ 单击 Next 按钮, 打开 Stage 1: Sample Size 对话框。在对话框中, 定义样本容量。它在 Units 的下拉式列表中通过选择 Counts 来定义抽样单位为总数。在 Values 下框中输入 6。表示在本阶段中选择 6 个整群。

⑥ 单击 Next 按钮, 进入 Stage 1: Output Variable 对话框, 见图 5-8。不做任何选择。

⑦单击 Next 按钮, 进入 Stage 1: Plan Summary 对话框, 见图 5-9。查看一下抽样设计。保持系统默认选择, 不再增加其他阶的抽样。

⑧单击 Next 按钮, 进入到 Stage 2: Selection Options 对话框, 见图 5-10。采用系统默认选择。

⑨单击 Next 按钮, 进入到 Stage 2: Output Files 对话框, 见图 5-11。选择 External file 选项, 输入 “E:\第 5 章 抽样方法\data05-16.sav”, 即将样本数据存在一个外部的数据文件中。由于选择了 PPS 抽样法, 因此, 在 Where do you want to save joint probabilities 选择项中, 必须指定联合概率的存储位置。在 file 后框中, 输入 “E:\第 5 章 抽样方法\data05-17.sav”, 即将联合概率存在一个外部的数据文件中。

⑩单击 Next 按钮, 进入到 Stage 2: Completion 对话框, 见图 5-12。保持系统默认选项, 去存储抽样方案文件, 并抽取一个样本。

⑪单击 Finish 按钮运行。

则在输出窗中出现抽样设计方案摘要表和第一阶段抽样摘要表, 见表 5-12 和表 5-13。

(1) 抽样设计方案摘要表, 见表 5-12。

表 5-12 抽样设计方案摘要

Summary			Stage 1
Design Variables	Cluster	1	县序号
Sample Information	Selection Method		PPS sampling without replacement
	Measure of Size		Obtained from data
	Minimum MOS Value		28
	Maximum MOS Value		69
	Number of Units Sampled		6
	Variables Created or Modified	Stagewise Inclusion (Selection) Probability	Inclusion Probability_1_
Analysis Information	Stagewise Cumulative Sample Weight		Sample Weight Cumulative_1_
	Estimator Assumption		Unequal probability sampling without replacement (using joint inclusion probabilities)
	Inclusion Probability		Obtained from variable Inclusion Probability_1_

Plan File: E:\2010年书稿\第5章_抽样方法\data5-06.csplan
Weight Variable: SampleWeight_Final_

表 5-12 回顾了整个抽样方案，表明抽样设计是一个一阶抽样，列出了在一阶中定义的群变量（县序号）及抽样方法（不重复 PPS 抽样）、规模测度（从数据中获取）、最小规模测度值（29）、最大规模测度值（68）、抽样单元的数量（6）、建立或修改的变量：阶段入样（选择）概率（Inclusion Probability_1_）、阶段累计抽样权重（Sample Weight Cumulative_1_）、分析信息：估计量假定（不等概不重复抽样）、入样概率（从入样概率变量中获取）。你可以用表中的信息来核实这是否就是你头脑中想要的设计的方案。

(2) 第一阶段抽样摘要表，见表 5-13。

表 5-13 第一阶段摘要

Summary for Stage 1

Number of Units Sampled		Proportion of Units Sampled	
Requested	Actual	Requested	Actual
6	6	33.3%	33.3%

表 5-13 列出了抽样第一阶段的摘要，第一列中显示抽样单元的数量，要求抽取的数量为 6、实际抽取的数量也为 6，第二列列出了层中抽样单元的比例，要求比例和实际比例相等均为 33.3%。

抽样结果：抽中的 6 个县的抽样数据资料被存放在“E:\第 5 章 抽样方法\data05-16.sav”中。打开该文件可见到显示结果。

联合概率则存放在“E:\第 5 章 抽样方法\data05-17.sav”的一个外部的数据文件中。

3.4.3 PPS 抽样的估计

1. 估值法

在有放回情况下，对总体总数 $Y = \sum_{i=1}^N Y_i$ 无偏估计为 $\hat{Y}_{\text{PPS}} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}$ ，该估计的均方偏差

$$\text{为 } V(\hat{Y}_{\text{PPS}}) = \frac{1}{n} \sum_{i=1}^N p_i \left(\frac{Y_i}{p_i} - Y \right)^2, \text{ 其无偏估计为 } v(\hat{Y}_{\text{PPS}}) = \frac{1}{n(n-1)} \left[\sum_{i=1}^n \left(\frac{y_i}{p_i} \right)^2 - n \hat{Y}_{\text{PPS}}^2 \right].$$

2. 实例分析

例 5.13 现对例 5.11 中抽中的 6 个县的奶牛的存栏头数进行调查，并将调查结果存放在“E:\第 5 章 抽样方法\data05-13a.sav”新建的存栏头数变量中，使对该地区的存栏头数的总数进行估计。

在 SPSS 中的解题步骤如下：

① 在数据编辑窗口中，打开 data05-13a.sav。

② 按 Analyze→Complex Samples→Descriptives 顺序，展开 Complex Samples Plan for Descriptives 对话框，见图 5-17。在 Plan 下方的 File 后面的输入框中输入“E:\第 5 章 抽样方法\data5-05.csplan”。在 Joint probabilities 中选择 Custom file，在其下的 file 后框中输入引号中的内容“E:第 5 章 抽样方法\data05-14.sav”。

③ 单击 Continue 按钮，进入如图 5-18 所示的描述统计对话框。

在左边变量名列表中，选择存栏头数并将其移入到 Measure 下框中，作为测量变量。

④ 单击 Statistics 按钮，展开如图 5-19 所示的描述统计选项卡。在 Summaries 中，关闭系统默认选项 Mean，选择 Sum 选项，在 Statistics 选项中，除默认选项以外，再选择 Confidence interval、Unweighted count、Population size、Design effect 选项。

⑤ 单击 Continue 按钮，返回 Complex Samples Descriptives 对话框。

⑥ 单击 OK 按钮，执行运算，在输出窗中得到输出结果。

⑦ 结果解释：

在输出窗中，产生一张表，见表 5-14。

表 5-14 对存栏头数变量计算的各个选定的统计量

Univariate statistics							
	Estimate	Standard Error	95%Confidence Interval		Design Effect	Population Size	Unweighted Count
			Lower	Upper			
Sean 存栏头数	97857	1336.486	94422	101293	0.016	21.915	6

表 5-14 中各列的含义同例 5.6 中的结果解释。由此可见用 PPS 抽样得到的 6 个县样本估计的该地区所养牛的总的存栏头数为 97857 头，标准误为 1336.486 头，该地区所养牛的总的存栏头数的 95%的置信区间为 (94422, 101293)。设计效应为 0.16。估计的总体容量为 21.915，所取的样本量为 6。

5.5 PPS Brewer 抽样

5.5.1 PPS Brewer 抽样概述

为使每一个总体单元的入样概率正比于规模测度 X_i ，在抽取第一个样本单元时，通过赋予总体中各单元一个修正概率，使总体中各单元的入样概率 $\pi_i (i=1, 2, \dots, N)$ 正比于规模测度 X_i ，这种抽样方法统称为 π PS 抽样。 π PS 抽样是抽样设计文献中涉及最多的抽样。

它属于不等概抽样方法中的一种。

PPS Brewer 抽样是不等概 π PS 抽样中的一种方法, 由 Brewer 于 1963 年提出。其基本想法是一次同时抽取两个样本单元, 使两个单元总的入样的概率正比于规模测度 X_i 。

为了做到这一点, Brewer 令 $p_i = \frac{X_i}{X}$ ($i=1, 2, \dots, N$), 其中 $X = \sum_{i=1}^N X_i$ 。以概率 p'_i 抽取第一个样本单元, $p'_i = \frac{p_i(1-p_i)}{1-2p_i}$, $i=1, 2, \dots, N$, 其中

$$D = \sum_{i=1}^N \frac{p_i(1-p_i)}{1-2p_i} = \frac{1}{2} \sum_{i=1}^N \frac{p_i(2-2p_i)}{1-2p_i} = \frac{1}{2} \left(1 + \sum_{i=1}^N \frac{p_i}{1-2p_i} \right)$$

取出第一个样本单元后不放回。当第一个样本单元为个体 U_j 时, 以概率 p''_i 抽取第二个样本单元, $p''_i = \frac{p_i}{1-p_j}$, $i \neq j$ 。对这一抽样, 可以证明, 总体的个体单元 U_i 的入样概率为 $\pi_i = 2p_i$, 而总体中两个个体 U_i 、 U_j 同时入样的概率

$$\pi_{ij} = \frac{p_i p_j}{D} \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j} \right) = \frac{2p_i p_j}{D} \frac{(1-p_i-p_j)}{(1-2p_i)(1-2p_j)}$$

因此, Brewer 法的实质是先设计好第一次抽样的概率, 第二次抽样的概率与 p_i 成正比, 使总的入样概率正比于 X_i 。

5.5.2 PPS Brewer 抽样在 SPSS 中实现

例 5.14 某市区有 7 个区, 各区去年底和今年现有的销售药品的商店数量列于表 5-15 中, 使用 PPS Brewer 法抽取两个区。

表 5-15 7 个区去年底和今年现有的销售药品的商店数

区序号	去年底商店数	现有商店数
1	33	33
2	36	39
3	71	63
4	48	51
5	19	23
6	69	77
7	48	45

在 SPSS 中，用 PPS Brewer 法进行随机抽样的步骤如下：

① 在 SPSS 中数据编辑窗中，建立变量区序号、去年底商店数和现有商店数三个变量，并将表 5-15 中的数据依次录入到这两个变量中，并将结果存放在 data05-18.sav 中。

② 单击 Analyze→Complex Samples→Select a Sample 顺序，打开如图 5-2 所示的 Welcome to Sampling Wizard 对话框。在 Sampling Wizard 对话框中，选择 Design a sample 选项，将插入点定位在其后 File 的文本编辑框中，输入“E:\第 5 章抽样方法\data5-07.csplan”。定义抽样方案文件名。

③ 单击 Next 按钮，进入 Stage 1: Design Variables 对话框。在中间的变量名列表中选择区序号，单击右侧第二个右移箭头，将区序号移入到 Cluster 下框中，定义区序号为第一阶群变量。

④ 单击 Next 按钮，进入 Stage 1: Sampling Method 对话框。在 Method 的 Type 下拉式选项中选择 PPS Brewer。在 Measure of Size(MOS)下，选择 Read from variable 选项，并将变量名列表中去年底商店数变量用右移箭头移到 Read from variable 的下框中，定义该变量为规模测度变量。

⑤ 单击 Next 按钮，进入 Stage 1: Output Variable 对话框，见图 5-8。不做任何选择。

⑥ 单击 Next 按钮，进入 Stage 1: Plan Summary 对话框，见图 5-9。查看一下抽样设计。保持系统默认选择，不再增加其他阶的抽样。

⑦ 单击 Next 按钮，进入到 Stage 2: Selection Options 对话框，见图 5-10。采用系统默认选择。

⑧ 单击 Next 按钮，进入到 Stage 2: Output Files 对话框，见图 5-11。选择 External file 选项，输入“E:\第 5 章 抽样方法\data05-19.sav”，即将样本数据存在一个外部的数据文件中。由于选择了 PPS 抽样法，因此，在 Where do you want to save joint probabilities 选择项中，必须做出联合概率存储在何处的决定。在 file 后框中，输入“E:\第 5 章 抽样方法\data05-20.sav”，即将联合概率存在一个外部的数据文件中。

⑨ 单击 Next 按钮，进入到 Stage 2: Completion 对话框，见图 5-12。保持系统默认选项，去存储抽样方案文件，并抽取一个样本。

⑩ 单击 Finish 按钮运行。

则在输出窗中出现抽样设计方案摘要表和第一阶段抽样摘要表，见表 5-16 和表 5-17。

(1) 抽样设计方案摘要表，见表 5-16。

表 5-16 回顾了整个抽样方案，表明抽样设计是一个一阶抽样，列出了在一阶中定义的群变量（区序号）及抽样方法（PPS Brewer 抽样）、规模测度（从变量去年底商店数中获取）、建立或修改的变量：阶段入样（选择）概率（Inclusion Probability_1_）、阶段累计抽样权重（Sample Weight Cumulative_1_）、分析信息：估计量假定（不等概不重

表 5-16 抽样设计方案摘要

Summary				Stage 1
Design Variables	Cluster	1		区序号
Sample Information	Selection Method			Brewer sampling with PPS ^a
	Measure of Size			Obtained from variable 去年底商店数
	Variables Created or Modified	Stagewise Inclusion (Selection) Probability		Inclusion Probability_1_
		Stagewise Cumulative Sample Weight		Sample Weight Cumulative_1_
Analysis Information	Estimator Assumption			Unequal probability sampling without replacement (using joint inclusion probabilities)
	Inclusion Probability			Obtained from variable Inclusion Probability_1_

Plan File: E:\2010年书稿\第5章 抽样方法\data5-07.csplan
Weight Variable: SampleWeight_Final_

a. This method samples two units per stratum

复抽样)、入样概率(从入样概率变量 Inclusion Probability_1_中获取)。你可以用表中的信息来核实这是否就是你头脑中想要的设计的方案。

(2) 第一阶段抽样摘要表, 见表 5-17。

表 5-17 第一阶段摘要

Summary for Stage 1			
Number of Units Sampled		Proportion of Units Sampled	
Requested	Actual	Requested	Actual
2	2	28.6%	28.6%

Plan File:E:\2010 年书稿\第 5 章抽样方法\data5-07.cs Plan

表 5-17 列出了抽样第一阶段的摘要, 第一列中显示抽样单元的数量, 要求抽取的数量为 2、实际抽取的数量也为 2, 第二列列出了层中抽样单元的比例, 要求比例和实际比

例相等均为 28.6%。

抽样结果:抽中的 2 个区的抽样数据资料被存放在“E:\第 5 章抽样方法\data05-19.sav”中。打开该文件可见到显示结果。

联合概率则存放在“E:\第 5 章 抽样方法\data05-20.sav”的一个外部的数据文件中。

5.5.3 PPS Brewer 抽样的估计

1. 估计方法

对于 π PS 抽样,一般都用由 Horvite-Thompson 提出的 HT 估计,即用 $\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i}$ 来估计总体总数 Y 。其中, π_i 是与样本单元 y_i 对应的入样概率。在 PPS Brewer 抽样中,

$n = 2$ 。总体总数的均方偏差为: $V(\hat{Y}_{HT}) = \sum_{i=1}^N \frac{(1-\pi_i)Y_i^2}{\pi_i} + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} Y_i Y_j$ 。而其用得

较多的一个无偏估计量为 $v(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{j>i}^n \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$ 。

2. 实例分析

例 5.15 现用例 5.14 中用 PPS Brewer 法抽取的存放在“E:\第 5 章 抽样方法\data05-19.sav”中的两个区现有商店数资料估计某市现在的商店总数。

在 SPSS 中的解题步骤如下:

① 在数据编辑窗口中,打开 data05-19.sav。

② 按 Analyze→Complex Samples→Descriptives 顺序,展开 Complex Samples Plan for Descriptives 对话框,见图 5-17。在 Plan 下方的 File 后面的输入框中输入“E:\第 5 章 抽样方法\data5-07.csplan”。在 Joint probabilities 中选择 Custom file,在其下的 file 后框中输入引号中的内容“E:\第 5 章 抽样方法\data05-20.sav”。

③ 单击 Continue 按钮,进入如图 5-18 所示的描述统计对话框。

在左边变量名列表中,选择现有商店数并将其移入到 Measure 下框中,作为测量变量。

④ 单击 Statistics 按钮,展开如图 5-19 所示的描述统计对话框。在 Summaries 中,关闭系统默认选项 Mean,选择 Sum 选项,在 Statistics 选项中,除默认选项以外,再选择 Confidence interval、Unweighted count、Population size、Design effect 选项。

⑤ 单击 Continue 按钮,返回 Complex Samples Descriptives 对话框。

⑥ 单击 OK 按钮,执行运算,在输出窗中得到输出结果。

⑦ 结果解释:

在输出窗中,产生一张表,见表 5-18。

表 5-18 对现有商店数变量计算的各个选定的统计量

Univariate statistics							
	Estimate	Standard Error	95%Confidence Interval		Design Effect	Population Size	Unweighted Count
			Lower	Upper			
Sun 现有商店数	333	22.261	50	616	0.094	5.723	2

表 5-18 中各列的含义同例 5.6 中的结果解释。由此可见用 PPS Brewer 抽样得到的 2 个区样本估计的该市现有商店数为 333 家，标准误为 22.261 家，该市现有商店数 95% 的置信区间为 (50, 616)。设计效应为 0.094。估计的总体容量为 5.723，所取的样本量为 2。

5.6 分层随机抽样

分层抽样 (Stratification) 是将总体分成若干个不重叠的小的总体，每个小总体称为一个层。分层抽样是在总体不重叠的小总体或层 (strata) 中挑选独立样本。例如，层可以是各个社会经济学组、工作类别、年龄组或种族。使用分层抽样的方法，可以保证重要子群的样本容量，改进全部估计的精度，以及在层与层之间使用不同的抽样方法。

5.6.1 样本容量的确定

5.6.1.1 分层样本容量的确定方法

在给定总的样本含量 n 时，常用的分层抽样各层样本含量的分配方法有以下几种：

1. 等额样本

每层取样本含量相等都为 $n_i = \frac{n}{K}$ 。其中 K 为所分的层数。分层随机抽样被称为**不成**

比例的 (disproportional)。

2. 按比例分配

样本含量按总体中各层个体单元的数量所占的比例分配， $n_i = n \cdot \frac{N_i}{N}$ 。当各层的个体单元数量 N_i 已知，而其他信息很少时，通常采用这种分配方案。此时的分层随机抽样被称为**成比例的** (proportional)。

3. 奈曼 (Neyman) 最优分配

分层抽样中， $n = \sum_{i=1}^K n_i$ 固定，使 $V(\bar{y}_{st}) = \sum_{i=1}^K W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_i^2$ 达到最小的样本含量分

$$\text{配为: } n_i = n \frac{W_i S_i}{\sum_{j=1}^K W_j S_j} = n \frac{N_i S_i}{\sum_{j=1}^K N_j S_j}, i=1, \dots, K$$

其中, $W_i = \frac{N_i}{N}$, N 为总体容量, S_i 为第 i 层样本的标准差。

5.6.1.2 实例分析

例 5.16 假设某地区 2001 年共有 904 家出口企业, 根据以往出口的金额, 我们将其分成五类, 见表 5-19, 现要求分层随机抽取 200 家进行调查, 各层如何分配抽样单元数?

表 5-19 某地区 904 家出口企业的分布情况

类别	以往出口额 (美元: 万)	企业数量	平均出口额 (美元: 万)	方差
1	2600~5000	40	3503.4	356554.26
2	1400~2600	74	1895.78	119535.6
3	700~1400	115	1036.38	34141.69
4	250~700	226	419.31	16565.56
5	50~250	449	123.26	3110.05

各层的抽样单元数取决于分配方法。

首先将表 5-19 中的数据在 SPSS 数据编辑窗口中建成数据文件, 见 data05-21.sav, 并使其处在打开状态。

第一种: 按等额样本分配

此法比较方便。

(1) 按 Transform→Compute Variable 顺序, 展开 Compute Variable 对话框。

(2) 在 Target Variable 下框中输入: 等额样本含量, 在 Numeric Expression 下框中输入以下双引号中的表达式: “rnd(200/5)”。

(3) 单击 OK 按钮, 在数据文件中增加一个新变量: 等额样本含量, 以及其计算结果的值从上到下各层依次为 40、40、40、40、40。

第二种: 按比例分配法

这种分配常常可以获得精度很好的估计。

(1) 按 Transform→Compute Variable 顺序, 展开 Compute Variable 对话框。

(2) 在 Target Variable 下框中输入: 按比例样本含量, 在 Numeric Expression 下框中输入以下双引号中的表达式: “rnd(200*企业数/904)”。

(3) 单击 OK 按钮, 在数据文件中增加一个新变量: 按比例样本含量, 以及其计算

结果的值从上到下各层依次为 9、17、26、51、200。其总数为 203，之所以比 200 大是由于取整进位的缘故造成的。

第三种：奈曼（Neyman）最优分配

使用本法估计各层样本含量的前提条件是要知道各层的方差或标准差的粗略估计值。

(1) 按 Transform→Compute Variable 顺序，展开 Compute Variable 对话框。

(2) 在 Target Variable 下框中输入：加权标准差，在 Numeric Expression 下框中输入以下双引号中的表达式：“企业数*Sqrt(方差)”。

(3) 单击 OK 按钮，在数据文件中增加一个新变量：加权标准差以及其计算结果。

(4) 以下求加权标准差和。按 Analyze → Descriptive Statistic → Descriptives 顺序，展开 Descriptives 对话框，从其左侧变量名源框中，选择加权标准差变量，移入到 Variable[s]下框中，单击 OK 按钮，则在输出窗中，出现输出结果，见表 5-20。

表 5-20 加权标准差和

Descriptive statistics			
	N	Sum	Std.Deviation
加权标准差	5	1.25E5	2844.38152
Valid N(listwise)	5		

由此可得加权标准差和为 1.25×10^5 。

(5) 按 Transform→Compute Variable 顺序，展开 Compute Variable 对话框。

(6) 在 Target Variable 下框中输入最优分配，在 Numeric Expression 下框中输入以下双引号中的表达式：“rnd(200*加权标准差/1.25E5)”。rnd()函数返回一个四舍五入的整数。

(7) 单击 OK 按钮，在数据文件中增加一个新变量：最优分配及其计算结果。

三种方法下各层的样本含量的分配见图 5-24。

上述计算后的结果连同 data05-21.sav 中的数据，已另存为 data05-21a.sav 中。

	类别	出口额上限	企业数	平均出口额	方差	等额样本含量	按比例样本含量	加权标准差	最优分配
1	1.00	5000.00	40.00	3503.40	356554.26	40.00	9.00	23884.87	38.00
2	2.00	2600.00	74.00	1895.78	119535.60	40.00	17.00	25584.70	41.00
3	3.00	1400.00	115.00	1036.38	34141.69	40.00	26.00	21249.09	34.00
4	4.00	700.00	226.00	419.31	16565.56	40.00	51.00	29087.84	47.00
5	5.00	250.00	449.00	123.26	3110.05	40.00	100.00	25039.75	40.00

图 5-24 三种方法的分层样本含量的计算结果

5.6.2 分层随机抽样过程

下面用实例的形式来说明在 SPSS 中实现分层随机抽样过程。

例 5.17 假设某地区 2001 年共有 904 家出口企业，根据以往出口的金额，我们将其分成五类，见表 5-19，现要求分层随机抽取 200 家进行调查，按奈曼（Neyman）最优分配法，在 SPSS 中应如何抽样？

在 SPSS 中的操作步骤如下：

① 在 SPSS 中数据编辑窗中，先建立层次分层变量和用以对应 904 家出口企业编号的编号变量的抽样框数据文件，见 data05-22.sav。

注：分层变量必须定义值标签，否则会在第 5 步中出错。

② 单击 Analyze→Complex Samples→Select a Sample 顺序，打开 Welcome to Sampling Wizard 对话框，见图 5-2。在 Sampling Wizard 对话框中，选择 Design a sample 选项，将插入点定位在其后 File 的文本编辑框中，输入“E:\第 5 章 抽样方法\data5-08.csplan”。定义抽样方案文件名。

③ 单击 Next 按钮，进入 Stage 1: Design Variables 对话框。在中间的变量名列表中选择出口金额分类，单击右侧第一个右移箭头，将出口金额分类移入到 Stratify By 下框中，定义层次为第一阶层变量；选择编号，单击右侧第二个右移箭头，将编号移入到 Cluster 下框中，定义编号为第一阶群变量。

④ 单击 Next 按钮，进入 Stage 1: Sampling Method 对话框。在本对话框中，不做任何选择，采用系统默认的 Simple Random Sampling（简单随机抽样法）。

⑤ 单击 Next 按钮，打开 Stage 1: Sample Size 对话框。在本对话框中，定义样本容量。它在 Units 的下拉式列表中通过选择 Counts 来定义抽样单位为总数。选择 Unequal values for strata 选项，表示各层选择不同的样品数量，单击激活的 Define 按钮，展开 Define Unequal Sizes 对话框，见图 5-7。单击第一层后面单元格，输入 38，单击第二层后面单元格，输入 41，单击第三层后面单元格，输入 34，单击第四层后面单元格，输入 47，单击第五层后面单元格，输入 40。这样完成对各层所要抽取的样品数量的录入。

⑥ 单击 Continue 按钮，返回 Stage 1: Sample Size 对话框。

⑦ 单击 Next 按钮，进入 Stage 1: Output Variable 对话框，见图 5-8。在输出变量对话框中不作任何选择。

⑧ 单击 Next 按钮，进入 Stage 1: Plan Summary 对话框，见图 5-9。查看一下抽样设计。保持系统默认选择，不再增加其他阶的抽样。

⑨ 单击 Next 按钮，进入到 Stage 2: Selection Options 对话框，见图 5-10。保持系统默认选择。

⑩ 单击 Next 按钮，进入到 Stage 2: Output Files 对话框，见图 5-11。选择 External file: 选项，输入“E:\第 5 章 抽样方法\data05-23.sav”，即将样本数据存在一个外部的数据文件中。

⑪ 单击 Next 按钮，进入到 Stage 2: Completion 对话框，见图 5-12。保持系统默认选项，表示你想存储抽样方案文件，并抽取一个样本。

⑫ 单击 Finish 按钮运行。

则在输出窗中出现抽样设计方案摘要表、第一阶矩阵说明表和第一阶段抽样摘要表，见表 5-21、表 5-22 和表 5-23。

(1) 抽样设计方案摘要表, 见表 5-21。

表 5-21 抽样设计方案摘要

Summary			Stage 1
Design Variables	Stratification	1	出口金额分类
	Cluster	1	编号
Sample Information	Selection Method		Simple random sampling without replacement
	Number of Units Sampled		Obtained from matrix specification
	Variables Created or Modified	Stagewise Inclusion (Selection) Probability	Inclusion Probability_1_
		Stagewise Cumulative Sample Weight	Sample Weight Cumulative_1_
Analysis Information	Estimator Assumption		Equal probability sampling without replacement
	Inclusion Probability		Obtained from variable inclusion Probability_1_

Plan File: E:\2010年书稿\第5章_抽样方法\data5-08.csplan
Weight Variable: SampleWeight_Final_

表 5-21 回顾了整个抽样方案, 表明抽样设计是一个一阶抽样, 列出了在一阶中定义的层变量(层次)、群变量(编号)及抽样方法(不重复简单随机抽样)、抽样单元数量(从矩阵说明中获取)、建立或修改的变量; 阶段入样(选择)概率(Inclusion Probability_1_)、阶段累计抽样权重(Sample Weight Cumulative_1_)、分析信息: 估计量假定(不重复等概抽样)、入样概率(从入样概率变量 Inclusion Probability_1_中获取)。你可以用表中的信息来核实这是否就是你头脑中想要的设计的方案。

(2) 第一阶矩阵说明表, 见表 5-22。

表 5-22 列出了为各层输入的不相等的抽样单元的数量。

(3) 第一阶段抽样摘要表, 见表 5-23。

表 5-22 第一阶矩阵说明表 Stage1

层次	Number of Units Sampled
第一层	38
第二层	41
第三层	34
第四层	47
第五层	40

表 5-23 第一阶段摘要

Summary for Stage 1

出口金额分类	Number of Units Sampled		Proportion of Units sampled	
	Requested	Actual	Requested	Actual
第一层	38	38	95.0%	95.0%
第二层	41	41	55.4%	55.4%
第三层	34	34	29.6%	29.6%
第四层	47	47	20.8%	20.8%
第五层	40	40	8.9%	8.9%

表 5-23 列出了抽样第一阶段的摘要，第一列列出了各层的名称，第二列显示各层抽样单元的数量（要求抽取的数量和实际抽取的数相同，分别为 38、41、34、47 和 40）第三列列出了各层中抽样单元的比例，要求比例和实际比例相等分别为 95.0%、55.4%、29.6%、20.8% 和 8.9%。

抽样结果：抽中的 200 个抽样数据资料被存放在“E:\第 5 章 抽样方法\data05-23.sav”中。打开该文件可见到显示结果。

5.6.3 分层随机抽样的估计

5.6.3.1 估计方法

如果分层抽样样本是从每一层独立抽取的，且每一层 \bar{Y}_i 有无偏估计 $\hat{\bar{Y}}_i$ ，则估计量 $\hat{\bar{Y}}_{st} = \sum_{i=1}^K W_i \hat{\bar{Y}}_i$ 是 \bar{Y} 的无偏估计量，其均方偏差为 $V(\hat{\bar{Y}}_{st}) = \sum_{i=1}^K W_i^2 V(\hat{\bar{Y}}_i)$ 。

(1) 当各层独立抽取的都是简单随机样本，且每层的 \bar{Y}_i 用简单估值时，则估计量 $\bar{y}_{st} = \sum_{i=1}^K W_i \bar{y}_i$ 是 \bar{Y} 的无偏估计量，其对应的均方偏差为 $V(\bar{y}_{st}) = \sum_{i=1}^K W_i^2 \frac{1}{n_i} (1-f_i) S_i^2$ ， $V(\bar{y}_{st})$ 的一个无偏估计为 $v(\bar{y}_{st}) = \sum_{i=1}^K W_i^2 \frac{1}{n_i} (1-f_i) s_i^2$ 。

(2) 当各层独立抽取的都是简单随机样本，且每层的样本量 n_i 足够大，用比估计法时， $\bar{y}_{RS} = \sum_{i=1}^K W_i r_i \bar{X}_i$ 是 \bar{Y} 的近似无偏估计，其均方偏差近似为

$$\begin{aligned} V(\bar{y}_{RS}) &= \sum_{i=1}^K W_i^2 \frac{1-f_i}{n_i} \frac{1}{N_i-1} \sum_{j=1}^{N_i} (Y_{ij} - R_i X_{ij})^2 \\ &= \sum_{i=1}^K W_i^2 \frac{1-f_i}{n_i} [S_{Y_i}^2 + R_i^2 S_{X_i}^2 - 2R_i \rho_i S_{Y_i} S_{X_i}] \end{aligned}$$

$V(\bar{y}_{RS})$ 的一个近似无偏估计为

$$\begin{aligned} v(\bar{y}_{RS}) &= \sum_{i=1}^K W_i^2 \frac{1-f_i}{n_i} \frac{1}{n_i-1} \sum_{j=1}^{n_i} (y_{ij} - r_i x_{ij})^2 \\ &= \sum_{i=1}^K W_i^2 \frac{1-f_i}{n_i} [s_{Y_i}^2 + r_i^2 s_{X_i}^2 - 2r_i s_{Y_i} s_{X_i}] \end{aligned}$$

其中 $\{X_{ij}\}$ ($i=1, \dots, K, j=1, \dots, N_i$) 为总体各单元的辅助指标，

$\{x_{ij}\}$ ($i=1, \dots, K, j=1, \dots, n_i$) 为各层样本单元的辅助指标,

$$W_i = \frac{N_i}{N}, \quad f_i = \frac{n_i}{N_i}$$

$$\bar{X}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}, \quad R_i = \frac{\bar{Y}_i}{\bar{X}_i}, \quad S_{Y_i}^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2$$

$$S_{X_i}^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2, \quad S_{XY_i} = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)(X_{ij} - \bar{X}_i)$$

$$\rho_i = \frac{S_{XY_i}}{S_{X_i} S_{Y_i}}, \quad \bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \quad r_i = \frac{\bar{y}_i}{\bar{x}_i}$$

$$s_{y_i}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2, \quad s_{x_i}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$$s_{xy_i} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(x_{ij} - \bar{x}_i)$$

分别为各层中的相应的量。

(3) 在有辅助指标 X 可用于估值分析时, 还可分别对指标量 Y 和辅助量 X 分别作分层简单估计, 再用比估值获得指标量的组合比估计, $\bar{y}_{RC} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X} = r_C \bar{X}$, 其中 \bar{X} 为辅助指标 X 的总体均值。当分层抽样的样本分配是按比例分配或最优分配时, \bar{y}_{RC} 是 \bar{Y} 的近似无偏估计, \bar{y}_{st} 估 \bar{Y} 的均方偏差近似地为

$$V(\bar{y}_{RC}) = E(\bar{y}_{st} - R\bar{x}_{st})^2 = \sum_{i=1}^K W_i^2 \frac{1-f_i}{n_i} \left([S_{Y_i}^2 + R^2 S_{X_i}^2 - 2R\rho_i S_{Y_i} S_{X_i}] \right)$$

其估计量为

$$v(\bar{y}_{RC}) = \sum_{i=1}^K W_i^2 \frac{1-f_i}{n_i} \left([s_{y_i}^2 + r_C^2 s_{x_i}^2 - 2r_C s_{xy_i}] \right)$$

式中符号的含义同上。

大样本量时使用比估计, 而层数较多, 各层样本量较小时, 推荐使用组合比估计。

5.6.3.2 实例分析

例 5.18 现对例 5.17 中的某地区 2001 年的 904 家出口企业, 按奈曼 (Neyman) 最优分配法, 用 data5-08.csplan 的抽样设计方案, 对分层随机抽取的 200 家进行了出口值调

查，获取的数据资料存放 data5-23a.sav 中。使其对 904 家总体的平均出口值进行估计。

在 SPSS 中的操作步骤如下：

① 在数据编辑窗口中，打开 Data05-23a.sav。

② 按 Analyze→Complex Samples→Descriptives 顺序，展开 Complex Samples Plan for Descriptives 对话框，见图 5-17。

③ 在 Plan 下方的 File 后面的输入框中输入“E:\第 5 章 抽样方法\data5-08.csplan”。

④ 单击 Continue 按钮，进入如图 5-18 所示的描述统计对话框。

在左边变量名列表中，选择出口值并将其移入到 Measure 下框中，作为测量变量。

⑤ 单击 Statistics 按钮，展开如图 5-19 所示的描述统计对话框。

在 Statistics 选项中，除默认选项以外，再选择 Confidence interval、Unweighted count、Population size、Design effect 选项。其他保持系统默认选项。

⑥ 单击 Continue 按钮，返回 Complex Samples Descriptives 对话框。

⑦ 单击 OK 按钮，执行运算，在输出窗中得到输出结果。

⑧ 结果解释：

在输出窗中，产生一张表，表 5-24。

表 5-24 对出口值变量计算的各个选定的统计量

Univariate Statistics							
	Estimate	Standard Error	95%Confidence Interval		Design Effect	Population Size	Unweighted Count
			Lower	Upper			
Mean 出口值	6.0822E2	7.23790	593.9411	622.4903	0.019	904.000	200

表 5-24 中各列的含义同例 5.6 中的结果解释。由此可见用分层抽样得到的 200 家样本估计的 904 家总体的平均出口值为 608.22 万美元，标准误为 7.2379 万美元，904 家总体的平均出口值的 95%的置信区间为（593.9411，622.4903），单位：万美元。

5.7 整群抽样

5.7.1 整群抽样概述

实施简单随机抽样、系统随机抽样和分层随机抽样通常是困难的，因为它们需要一个完整的抽样框。例如，像抽样城市或抽样医院或抽样学校的名单清单是容易得到的，但获取抽样个人或家庭的名单清单非常困难。当一个完整的总体清单不可得到时，整群抽样（Cluster Sampling）是有用的。

所谓整群抽样就是在多阶抽样中, 当某一阶抽样单元被抽中, 该单元包含许多下一级时, 在被抽中的单元内不再进行下一级的抽样, 而是对该单元内的下一级单元进行全面调查。因此, 整群可以是学校、医院或地区, 而且样本单元可以是学生、病人或市民。整群抽样常用于多级设计和区域抽样中。

当各个群内包含的次级抽样单元个数比较相近时, 常采用简单随机抽样; 而当各个群的规模相差比较悬殊时, 则多采用 PPS 抽样。

5.7.2 整群抽样在 SPSS 中的实现

1. 当各个群内包含的次级抽样单元个数比较相近时, 常采用简单随机抽样或系统抽样。具体做法参见简单随机抽样和系统抽样。

2. 而当各个群的规模相差比较悬殊时, 则多采用 PPS 抽样。具体做法可参见 PPS 抽样, 当只抽取两个群时, 也可参见 PPS Brewer 抽样。

5.7.3 整群抽样的估计

1. 简单随机的整群抽样

总体总数 Y 的无偏估计为: $\hat{Y}_{\text{CSE}} = \frac{K}{k} \sum_{i=1}^k \sum_{j=1}^{N_{\theta_i}} Y_{\theta_{ij}}$ 。

\hat{Y}_{CSE} 的均方偏差为: $V(\hat{Y}_{\text{CSE}}) = \frac{K^2}{k} \left(1 - \frac{k}{K}\right) \frac{1}{K-1} \sum_{i=1}^K \left(\sum_{j=1}^{N_i} Y_{ij} - \frac{Y}{K} \right)^2$ 。其无偏估计为

$$v(\hat{Y}_{\text{CSE}}) = \frac{K^2}{k} \left(1 - \frac{k}{K}\right) \frac{1}{k-1} \sum_{i=1}^k \left(\sum_{j=1}^{N_i} Y_{\theta_{ij}} - \frac{\hat{Y}_{\text{CSE}}}{K} \right)^2$$

2. 有放回 PPS 整群抽样

总体总数 Y 的无偏估计为: $\hat{Y}_{\text{CSE}} = \frac{1}{k} \sum_{i=1}^k \frac{1}{p_{\theta_i}} \left(\sum_{j=1}^{N_{\theta_i}} Y_{\theta_{ij}} \right)$

\hat{Y}_{CSE} 的均方偏差为: $V(\hat{Y}_{\text{CSE}}) = \frac{1}{k} \sum_{i=1}^K p_i \left(\frac{1}{p_i} \sum_{j=1}^{N_i} Y_{ij} - Y \right)^2$, 其无偏估计为

$$v(\hat{Y}_{\text{CSE}}) = \frac{1}{k(k-1)} \sum_{i=1}^k \left(\frac{1}{p_{\theta_i}} \sum_{j=1}^{N_i} Y_{\theta_{ij}} - \hat{Y}_{\text{CSE}} \right)^2$$

5.7.4 整群抽样的实例分析

例 5.19 为调查一块 5000 平方米荒地上每平米蝗蛹的数量，将其人为地均分为 500 个地块，每块 10 平方米，现从 500 个地块中，用简单随机抽样的方法抽取 20 个作为一级样本单元，再对抽中的地块调查每一平米的蝗蛹数，依次估计整块荒地的蝗蛹数。

在 SPSS 中的操作步骤如下：

① 在 SPSS 中数据编辑窗中，建立数据文件，见 data05-24.sav。

② 单击 Analyze→Complex Samples→Select a Sample 顺序，打开如图 5-2 所示的 Welcome to Sampling Wizard 对话框。在 Sampling Wizard 对话框中，选择 Design a sample 选项，将插入点定位在其后 File 的文本编辑框中，输入“E:\第 5 章抽样方法\data5-09.csplan”。定义抽样方案文件名。

③ 单击 Next 按钮，进入 Stage 1: Design Variables 对话框。在中间的变量名列表中选择地块编号，单击右侧第二个右移箭头，将地块编号移入到 Cluster 框中，定义地块编号为第一阶群变量。

④ 单击 Next 按钮，进入 Stage 1: Sampling Method 对话框。在 Method 的 Type 下拉式选项中选择 Simple Random Sample（简单随机抽样法）。其他采用系统默认的 without replacement(WOR)选项，即在本阶段中，利用默认的不重复（WOR）简单随机抽样法，抽取的地块编号作为主要的抽样单元。

⑤ 单击 Next 按钮，打开 Stage 1: Sample Size 对话框。在本对话框中，定义样本容量。它在 Units 的下拉式列表中通过选择 Counts 来定义抽样单位为总数。在 Values 下框中输入 20。表示在本阶段中选择 20 个样本单元。

⑥ 单击 Next 按钮，进入 Stage 1: Output Variable 对话框，见图 5-8。不做任何选择。

⑦ 单击 Next 按钮，进入 Stage 1: Plan Summary 对话框，见图 5-9。查看一下抽样设计。保持系统默认选择，不再增加其他阶的抽样。

⑧ 单击 Next 按钮，进入到 Stage 2: Selection Options 对话框，见图 5-10。采用系统默认选择。

⑨ 单击 Next 按钮，进入到 Stage 2: Output Files 对话框，见图 5-11。选择 External file 选项，输入“E:\第 5 章 抽样方法\data05-25.sav”，即将样本数据存在一个外部的数据文件中。

⑩ 单击 Next 按钮，进入到 Stage 2: Completion 对话框，见图 5-12。保持系统默认选项，去存储抽样方案文件，并抽取一个样本。

⑪ 单击 Finish 按钮运行。

则在输出窗中出现抽样设计方案摘要表和第一阶段抽样摘要表，参见例 5.4 中对表 5-3 和表 5-4 的解释。

抽样数据被存放在“E:\第5章 抽样方法\data05-25.sav”中。

现在对抽中的20块地做整群调查，结果见表5-25。

表5-25 200块地10列20行的蝗蛹数统计表

列 行	1	2	3	4	5	6	7	8	9	10
1	6	4	4	21	23	14	26	19	20	16
2	24	26	9	14	5	20	22	24	29	24
3	18	17	23	9	29	24	27	17	15	26
4	14	7	20	24	15	37	19	12	38	28
5	24	24	8	39	28	23	32	34	37	39
6	25	18	23	37	22	23	16	22	33	37
7	32	19	25	23	30	21	31	23	21	25
8	35	12	12	22	16	15	12	36	34	27
9	29	21	39	16	13	19	6	20	9	1
10	2	3	9	7	23	14	11	20	16	7
11	22	17	26	16	27	22	30	34	29	23
12	33	17	40	24	17	18	15	16	21	27
13	15	10	4	6	8	17	13	6	11	12
14	17	18	12	11	10	6	22	14	4	27
15	13	9	5	7	9	17	14	9	4	14
16	18	23	13	15	8	16	24	26	28	26
17	33	5	26	30	11	10	15	17	10	21
18	26	15	13	17	3	12	1	5	5	7
19	7	32	4	6	9	4	3	11	9	12
20	15	1	1	6	5	5	6	7	2	1

⑫ 在 SPSS 中数据编辑窗中，打开 data05-25.sav，新建变量名蝗蛹数，并将表 5-25 中数据依次录入到蝗蛹数变量中，并将其另存为 data05-25a.sav。

⑬ 按 Analyze→Complex Samples→Descriptives 顺序，展开 Complex Samples Plan for Descriptives 对话框，见图 5-17。在 Plan 下方的 File 后面的输入框中输入“E:\第5章 抽样方法\data5-09.csplan”。

⑭ 单击 Continue 按钮，进入如图 5-18 所示的描述统计对话框。

在左边变量名列表中，选择蝗蛹数并将其移入到 Measure 下框中，作为测量变量。

⑮ 单击 Statistics 按钮，展开如图 5-19 所示的描述统计选项卡。

在 Summaries 中，关闭系统默认选项 Mean，选择 Sum 选项，在 Statistics 选项中，除默认选项以外，再选择 Confidence interval、Unweighted count、Population size、Design effect 选项。

⑯ 单击 Continue 按钮，返回 Complex Samples Descriptives 对话框。

⑰ 单击 OK 按钮，执行运算，在输出窗中得到输出结果。

⑱ 结果解释：

在输出窗中，产生一张表，表 5-26。

表 5-26 对蝗蛹数变量计算的各个选定的统计量

Univariate Statistics							
	Estimate	Standard Error	95% Confidence Interval		Design Effect	Population Size	Unweighted Count
			Lower	Upper			
Sun 蝗蛹数	8.78E4	4553.31921	78269.79	97330.21	1.826	5000.000	200

表 5-26 中各列的含义同例 5.6 中的结果解释。由此可见用简单随机整群抽样得到的 20 个群 200 个一平方米小块样本估计的总体蝗蛹总数为 87800 只，标准误为 4553.31921 只，有 95% 的把握认为蝗蛹总数将落在 78270 至 97330 之间。设计效应为 1.826。

5.8 多 阶 抽 样

5.8.1 多阶抽样概述

多阶抽样 (Multiple stages) 是将上述各种方法组合在一起所形成的一个多步骤的综合抽样方法，适用于大范围的抽样调查。

在多阶抽样中，先根据整群挑选第一阶样本。然后用从已挑选的整群中抽取子样本建立第二阶样本。如果第二阶样本是建立在子群上的，则你然后可为抽样增加第三阶。例如，在调查的第一阶段，抽签抽取一个城市样本。然后，从选择的城市中，抽样得到家庭。最后，从已选择的家庭中，抽签选择个体。在 SPSS 中，一般多阶抽样的最多分阶数为 3。

下面是一个在乡村实施的多阶抽样的例子：

- 把乡村（或人口普查片区）看作层，选择一定数量乡村的一个随机样本。
- 在每个选择的乡村内，抽取长方形区域的一个群随机样本。
- 在每个选择的区内，每隔第 10 个住宅抽取一个系统随机样本。
- 在每个选择的家庭家庭内，给样本随机选择一个成年人。

多阶样本常见于社会科学研究中。它们实施起来比简单随机抽样更简单，而且比整

群抽样的单个方法提供总体更广泛的抽样。

5.8.2 多阶抽样实例分析

例 5.20 为方便、有效地获取具有代表性的城市家庭数量方面的资料，我们使用复合采样方法，整个抽样设计结构要求如下：

阶段	层	群
1	大行政区	省
2	地区	城市
3	小区	

在第三阶段中，主要的抽样单元是家庭，且要对选择的家庭进行调查。但是，由于仅从城市层面而言可以很容易得到资料，所以，我们先执行设计的前两个阶段，然后收集有关城市小区和抽到的城市中家庭的数量方面资料。城市层面有用的资料已汇集在数据文件 data05-26.sav 中，注意，本数据文件里小区变量里的所有值都是 1。这是一个适合于“True”变量的占位符，在执行设计的前两个阶段后将收集到它的值。现在我们使用 Complex Samples Sampling Wizard 来指定完整的三阶抽样设计，然后用前两个阶段的抽样设计抽取样本。据此样本收集有关城市小区和抽到的城市中的家庭的数量方面资料，做成数据文件，用它来抽取本设计第三阶的样本。

操作步骤如下：

① 在 SPSS 数据编辑窗口中打开 Data05-26.sav，并单击 Analyze→Complex Samples→Select a Sample 顺序，打开 Sampling Wizard 对话框。

② 在 Sampling Wizard 对话框中，选择 Design a sample 选项，将插入点定位在其后 File 的文本编辑框中，输入“E:\第5章 复式采样\data5-10.csplan”。

③ 单击 Next 按钮，进入 Stage 1: Design Variables 对话框。

④ 在中间的变量名列表中选择大行政区，单击右侧第一个右移箭头，将大行政区移入到 Stratify by 下框中，定义大行政区为第一阶层变量。

⑤ 在中间的变量名列表中选择省，单击右侧第二个右移箭头，将省移入到 Cluster 下框中，定义省为第一阶群变量。

⑥ 单击 Next 按钮，进入 Sampling Method 对话框。在本对话框中，不做任何选择，即采用系统默认的选择。这个设计结构意味着要为各个区域抽取一个独立样本。在本阶段中，利用默认的简单随机抽样法，抽取的省作为主要的抽样单元。

⑦ 单击 Next 按钮，打开 Sample Size 对话框。在 Units 的下拉式列表中选择 Counts。在 Values 下框中输入 3。表示在本阶段中选择 3 个单元。

⑧ 单击 Next 按钮, 进入 Output Variable 对话框。在输出变量对话框中不做任何选择, 直接单击 Next 按钮, 进入 Summary 对话框, 选择 Yes, 单击左侧矩形框中的 Add Stage 2, 或单击 Next 按钮, 进入到 Stage 2: Design Variables 对话框。

⑨ 在中间的变量名列表中选择地区, 单击右侧第一个右移箭头, 将地区移入到 Stratify by 下框中, 定义地区为第二阶的层变量。

⑩ 在中间的变量名列表中选择城市, 单击右侧第二个右移箭头, 将城市移入到 Cluster 下框中, 定义城市为第二阶群变量。

⑪ 单击 Next 按钮, 进入 Stage 2: Sampling Method 对话框。采用系统默认选择。这个设计结构意味着要为各个地区抽取独立样本。在本阶段中, 利用默认的简单随机抽样法, 抽取的城市作为主要的抽样单元。

⑫ 单击 Next 按钮, 进入 Stage 2: Sample Size 对话框。在 Units 下拉式列表中选择 Proportions 选项, 在 Values 下框中输入 0.1, 作为从每层中抽样单元的比例。

⑬ 单击 Next 按钮, 进入 Stage 2: Output Variable 对话框。在输出变量对话框中不做任何选择, 直接单击 Next 按钮, 进入 Stage 2: Plan Summary 对话框, 选择 Yes, 单击左侧矩形框中的 Stage 3, 或单击 Next 按钮, 进入到 Stage 3: Design Variables 对话框。

⑭ 在中间的变量名列表中选择小区, 单击右侧第一个右移箭头, 将小区移入到 Stratify by 下框中, 定义小区为第三阶的层变量。

⑮ 单击 Next 按钮, 进入 Stage 3: Sampling Method 对话框。采用系统默认选择。这个设计结构意味着要为各个城市小区抽取独立样本。在本阶段中, 利用默认的简单随机抽样法, 抽取的家庭单元作为主要的抽样单元。

⑯ 单击 Next 按钮, 进入 Stage 3: Sample Size 对话框。在 Units: 下拉式列表中选择 Proportions 选项, 在 Values 下框中输入 0.2, 作为从每层中抽样单元的比例。

⑰ 单击 Next 按钮, 进入 Stage 3: Output Variable 对话框。在输出变量对话框中不做任何选择, 直接单击 Next 按钮, 进入 Stage 3: Plan Summary 对话框, 单击 Next 按钮, 进入到 Draw Sample: Selection Options 对话框。

⑱ 在 Do you want to draw a sample? 下, 选择 Yes 该项, 意味着要抽取一个样本。在 Stages 的下拉式列表中选择 1、2, 即要抽取前两个阶段的样本。

在 What type of seed value do you want to use 下, 选择 Custom Value, 输入 241972 作为随机种子值。

⑲ 单击 Next 按钮, 进入 Draw Sample: Output Files 对话框。选择 External file 选项, 输入 “E:\第 5 章 抽样方法\data05-27.sav”, 即将样本数据存在一个外部的数据文件中。单击 Next 按钮, 进入 Completing the Sampling Wizard 对话框。

⑳ 单击 Finish 按钮运行。

这些选择产生抽样方案文件 data5-10.csplan, 并根据抽样设计方案中的前两个阶段抽

取一个样本。

② 运行结果

在输出窗中，得到输出信息。因信息量较大，故单击输出信息的先后顺序分别列出并单独解释。

- 抽样设计方案摘要表，见表 5-27。

对该表的解释参照表 5-3 的解释。

表 5-27 抽样设计方案摘要

Summary			Stage 1	Stage 2	Stage 3
Design Variables	Stratification	1	大行政区	地区	小区
	Cluster	1	省	城市	
Sample Information	Selection Method		Simple random sampling without replacement	Simple random sampling without replacement	Simple random sampling without replacement
	Number of Units Sampled		3		
	Variables Created or Modified	Stagewise Inclusion (Selection) Probability	Inclusion (Selection) Probability for Stage 1	Inclusion (Selection) Probability for Stage 2	Inclusion Probability_3_
		Stagewise Cumulative Sample Weight	Cumulative Sampling Weight for Stage 1	Cumulative Sampling Weight for Stage 2	Sample Weight Cumulative_3_
Analysis Information	Proportion of Units Sampled			.1	.2
	Estimator Assumption		Equal probability sampling without replacement	Equal probability sampling without replacement	Equal probability sampling without replacement
	Inclusion Probability		Obtained from variable Inclusion (Selection) Probability for Stage 1	Obtained from variable Inclusion (Selection) Probability for Stage 2	Obtained from variable Inclusion Probability_3_

Plan File: E:\2010年书稿\第5章 抽样方法\data5-10.csplan
Weight Variable: Final Sampling Weight

- 第一阶段抽样摘要表，见表 5-28。

其解释参阅上题对表 5-4 的解释。

表 5-28 第一阶段抽样摘要表

Summary for Stage 1				
大行政区	Number of Units Sample		Proportion of Units Sampled	
	Requested	Actual	Requested	Actual
1	3	3	20.0%	20.0%
2	3	3	20.0%	20.0%
3	3	3	20.0%	20.0%
4	3	3	20.0%	20.0%

- 第二阶段抽样摘要表，见表 5-29。

表 5-29 第二阶段抽样摘要表

Summary for stage 2						
大行政区	省	地区	Numhber of Units Sampled		Proportion of Units Sampled	
			Requested	Actual	Requested	Actual
1	2	7	3	3	10.0%	10.0%
		8	3	3	10.0%	10.0%
		9	3	3	10.0%	10.0%
		10	3	3	10.0%	10.0%
		11	3	3	10.0%	10.0%
		12	3	3	10.0%	10.0%
	6	31	3	3	10.0%	10.0%
		32	3	3	10.0%	10.0%
		33	3	3	10.0%	10.0%
		34	3	3	10.0%	10.0%
		35	3	3	10.0%	10.0%
	11	36	3	3	10.0%	10.0%
		61	3	3	10.0%	10.0%
		62	3	3	10.0%	10.0%
		63	3	3	10.0%	10.0%
		64	3	3	10.0%	10.0%
2	17	65	3	3	10.0%	10.0%
		66	3	3	10.0%	10.0%
		97	3	3	10.0%	10.0%
		98	3	3	10.0%	10.0%
		99	3	3	10.0%	10.0%
		100	3	3	10.0%	10.0%
		101	3	3	10.0%	10.0%
		102	3	3	10.0%	10.0%
	18	103	3	3	10.0%	10.0%
		104	3	3	10.0%	10.0%
		105	3	3	10.0%	10.0%
		106	3	3	10.0%	10.0%

其解释参阅对表 5-4 的解释。

- 数据编辑窗口中的抽样结果

你可在正在工作的数据编辑窗口中，看到本次的抽样结果。

5 个新增的带下画线的变量被存放在工作的数据文件中。在 Inclusion probability_1_ 下存放的是第一阶段体现的内含概率值。在 Sample Weight Cumulative_1_ 下存放的是第一

阶段累积抽样权重在 Inclusion Probability_2_下存放的是第二阶段体现的内含概率值。在 Sample Weight Cumulative_2_下存放的是第二阶段累积抽样权重在 Sample Weight_Final_下存放的是最终的抽样权重。

现在我们对每个选取的城市进行调研，将已获得的城市小区和家庭单元资料存放在 Data5-27a.sav。使用这个文件和 Sampling Wizard 去抽取本设计第三阶的样本。

具体操作步骤如下：

① 在 SPSS 数据编辑窗口中打开 Data5-27a.sav，并单击 Analyze→Complex Samples →Select a Sample 顺序，打开 Sampling Wizard 对话框。

② 在 Sampling Wizard 对话框中，选择 Draw a sample 选项，将插入点定位在其后 File 的文本编辑框中，输入“E:\第5章 复式采样\data5-10.csplan”。

③ 单击 Next 按钮，进入 Plan Summary 对话框。

在 Which stages have already been sampled 下面 stages: 后面的下拉式列表的显示项中选择 1, 2 选项。表示已经对前两个阶段进行了抽样。

④ 单击 Next 按钮，进入 Draw Sample: Selection Options 对话框。

在 What type of seed value do you want to use 下，选择 Custom Value，输入 241972 作为随机种子值。

⑤ 单击 Next 按钮，进入 Draw Sample: Output Files 对话框。

选择 External file: 选项，输入“E:\第5章 抽样方法\data05-28.sav”，即将样本数据存在一个外部的数据文件中。

⑥ 单击 Next 按钮，进入 Completing the Sampling Wizard 对话框。

在 What do you want to do 下，选择第二个选项：Paste the syntax generated by the Wizard into a syntax window。

⑦ 单击 Finish 按钮运行。

打印本例抽样摘要，会产生一个能引起 Output Viewer 问题的笨重的表格。故在本例操作的第6步中，在 Completing the Sampling Wizard 对话框中，选择了 Paste the syntax generated by the Wizard into a syntax window。这可以给我们一个修改程序的机会。（为关闭抽样摘要显示，在 PRINT 子命令中用 CPS 替代 SELECTION。然后在 Syntax 窗口中单击 Run，在其下拉式菜单单击 All 运行程序。）

值得一提的是在 Syntax 窗口运行前，先要在 Syntax 窗口中对上面的程序的两个地方作修改：

在 /CRITERIA STAGES=SEED=241972 中，需要改成：

/CRITERIA STAGES=3 SEED=241972 （要求对第三阶段抽样）

在 /PRINT SELECTION 中，需要改成：

/PRINT CPS

则运行后，在数据文件 `data05-28.sav` 中，你可以看到抽样结果。三个新变量存储到该数据文件中，描述了第三阶抽样的入样概率、累计抽样权重及最终抽样权重。这些新权重考虑了前两个阶段抽样中的权重计算。

在输出窗口中，还出现如图 5-25 所示的结果。首先显示的是 CSSELECT 过程的程序清单。其次是本次做的任务名称为用复式采样选择一个样本。下面给出的是本次工作所用的数据是文件的名称。最后一张表中所显示的是工作数据文件中未加权样品数的有效数、无效数和总数。

```
* Sampling Wizard.
CSSELECT
/PLAN FILE='E:\2010年书稿\第5章 抽样方法\data5-10.csplan'
/CRITERIA STAGES=3 SEED=241972
/CLASSMISSING EXCLUDE
/DATA RENAMEVARS
/SAMPLEFILE OUTFILE='E:\2010年书稿\第5章 抽样方法\data05-28.sav'
/PRINT CPS.
```

Complex Samples: Selection

[DataSet1] E:\2010年书稿\第5章 抽样方法\data05-27a.sav

Case Processing Summary

	Unweighted Cases
Valid	30112
Invalid	0
Total	30112

图 5-25 显示在输出窗口中的信息

第6章 假设检验

假设检验 (Hypothesis Testing) 是推断统计中的另一类重要问题。由于受到总体的无限性或实际研究过程中各种条件的有限性的限制, 在实际的研究中, 我们通常采用从总体中随机取样的方法, 用样本提供的信息来判断总体是否服从某种分布或总体是否具有某种指定的特征。在这些情况下, 必须借助于统计学提供的假设检验的方法来加以判定。

本章将主要介绍假设检验的基本思想和参数性假设检验方法。

6.1 假设检验概述

6.1.1 何谓统计假设

关于“假设”一词, 在新华词典中, 也将其称为假说。是指在科学研究中用来解释某种有待证明的论题的说明。假设被充分证明后就是理论。

在统计学中, 所谓的**统计假设**是指总体的分布类型, 或关于参数, 即总体的某个统计测度数的假定。

由此可知, 并非每个假设都是统计假设。

例如:

- ① 火星上有人居住;
- ② 天下乌鸦一般黑。

它们是用来表达判断的句子, 即命题。

证明这些命题是真是假, 是很简单的, 只要能找到一个来自地球以外的人在火星上居住, 命题 1 就成立, 或只要找到一只乌鸦是白色的或非黑色的, 就可证明命题 2 是假的。

由于在上述的两个例子中不涉及总体的分布类型, 也不涉及总体的分布特征, 即参数, 所以, 它们是一般意义上的假设, 而不是特指的统计假设。

如果把 2 稍加修改“天下乌鸦 99% 是黑色的”, 则它就是统计假设, 因为它给出了一个统计参数, 即 99% 的百分比的假定。

6.1.2 可否直接根据试验结果数据值大小来做出拒绝或不拒绝统计假设的结论

对于统计假设, 一般我们只有两种选择, 一种是拒绝它, 另一种是不拒绝它。有一部分研究者, 喜欢直接用试验结果的数值大小来直接给出拒绝或不拒绝统计假设的统计

结论。他们把统计假设当作假设或命题来对待了。事实上，单凭试验数据表面值的大小是不能直接作为拒绝或不拒绝统计假设的依据的。

例 6.1 掷一颗六面体的骰子 300 次，结果见表 6-1。根据表 6-1 的实验数据，可否认为它是不均匀的？

表 6-1 300 次试验结果分布表

点数 i	1	2	3	4	5	6
频数 i	43	49	56	45	66	41

例 6.2 把 20 名原始成绩基本相同，各方面条件都很相似的自行车运动员随机分成两组，两组自行车运动员在比赛期间采用两种不同的力量训练方法进行训练，第一组把力量训练的全部练习安排在整个比赛季节进行，而第二组把力量训练的全部练习安排在整个比赛季节的后半期进行，在前半期根本不用力量练习，两组训练方法的效果用从原地出发的 500 米骑行成绩的提高的多少来评定。（单位：秒）。实验测得的数据资料见表 6-2。

表 6-2 两种力量训练后 500 米骑行成绩的提高值统计结果

统计量	\bar{x}	s	n
第一组	1.42	0.5412	10
第二组	1.02	0.3853	10

由表 6-2 可知，第一组运动员运动成绩提高的均数要比第二组高 0.4 秒。由题意可知，两组原始成绩相同，身体各方面的条件基本相似，而且分组又是随机的，我们能否用均数之间的差异本身来说明第一组的训练方法比第二组的训练方法的效果更好呢？

例 6.3 在疫苗是否有效的实验中，将 401974 名志愿者随机地分成两组，一组 200745 人为处理组，另一组 201229 人为对照组。对处理组的人接种疫苗，对对照组的人接种不含疫苗的安慰剂。实验结果见表 6-3。

表 6-3 处理组与对照组的发病率情况表

	人数	病例数	发病率
处理组	200745	57	0.000284
对照组	201229	142	0.000706

从发病率的百分比来看，前者为 0.0284%，后者为 0.0706%，是否认为接种疫苗后发病率明显低于不接种疫苗的对照组呢？

例 6.4 从自动车床加工中轴的成品中随机抽取 11 根，测得它们的直径（单位：mm）如下：

10.52, 10.41, 10.32, 10.18, 10.64, 10.77, 10.82, 10.67, 10.59, 10.38, 10.49

问这批零件的直径 X 服从正态分布吗？

在上面的例子中，有两个共同的特点：

1. 都有一个统计假设

在例 6.1 中，通过所做的 300 次试验的结果，可否认为“它是不均匀的”？这个问题同要拒绝“骰子每个点出现的概率都是 $1/6$ ”这样一个统计假设是等价的。

在例 6.2 中，“第一组的训练方法比第二组的训练方法的效果更好”同要检验参数“ $\bar{x}_1 \geq \bar{x}_2$ ”的统计假设是等价的。

在例 6.3 中，“接种疫苗后发病率明显低于不接种疫苗的对照组”同要检验参数“ $p_1 \leq p_2$ ”的统计假设是等价的。

在例 6.4 中，“零件的直径 X 服从正态分布”同“ $X \sim N(\mu, \sigma^2)$ ”的统计假设是等价的。

2. 都要借助于抽样数据

另一个共同的特点，就是都要通过实际测得的数据资料来回答题中的问题。

我们知道，只要对某个观察特征进行测试，则得到的数据资料并不总是相等，而是参差不齐，具有波动性的。引起数据资料差异的原因有很多，但从统计学的角度而言，主要是由于各种误差的存在。

统计误差通常指测量值与其真值之差。如，某人身高的真值为 211cm，通过对其测试得到的测量值为 211.1 cm，则其统计误差为 0.1cm。而把测量值与其样本均值之差称偏差，又称离均差。如一名跳远运动员在一次训练课中，在状态最佳的三次跳远测试中，测得其跳远成绩分别为：6.56m、6.68m、6.62m，则其三次跳远的样本均值为：6.62m，偏差分别为：-0.06m、0.06m、0.00m。总偏差和恒为 0。

统计误差是以下几种误差的总称，也即它主要来源于以下几个途径：

(1) 系统误差：它是指在收集数据资料的过程中，由于测量仪器的不准或测试人员掌握测试标准的尺度偏高或偏低等原因造成的使测试结果呈倾向性的偏大或偏小的误差。

(2) 过失误差：它是指在测试、记录、抄写或录入计算机过程中，由于工作人员的粗心大意所造成的误差。该误差大小不一且没有方向性，它的存在可使分析结果与实际出现严重背离。

(3) 随机测量误差：指在观测过程中由于各种偶然因素造成的对同一观测对象的多次测量得到不完全相同的测量结果的没有固定倾向的或高或低的误差。这种误差是不可避免的，但要采取措施尽量减少。必要时应对随机误差进行分析。

(4) 抽样误差：顾名思义，是由抽样引起的样本和总体之间的差异。由于总体中的个体存在差异，即使进行随机化抽样，多次抽样得到多个样本，其均值也会有所不同。抽样误差是由于个体之间的差异造成的。这些个体间的差异同时反应了样本与总体的差异。个体差异是无法消除的，但可以控制在一个可以接受的范围中，即研究的精度控制

的范围中。统计学中的统计推断，就是在根据样本特征推断总体特征时研究抽样误差的影响。

通常，样本越大，则抽样误差越小，样本特征与总体特征越接近；而样本越小，抽样误差越大。虽然在实际工作中，受人力、物力、时间限制，样本不可能太大，因此在使用样本特征推断总体特征时要谨慎。

(5) 条件误差：对被试对象施加所要研究的条件刺激后所引起的在去除系统误差、过失误差、随机测量误差和抽样误差后的测量值与正常情形下同一测试指标值之间的差异称条件误差。

上述误差，并不总是都存在于总的误差中的。例如，系统误差是可控误差，可以在测试前通过对观测人员进行统一尺度的培训、校准测试仪器、统一使用标准的试剂等方法来避免系统误差的出现，以增强原始资料的准确性。过失误差也是可控误差，只要加强工作人员的责任心，它也是可以避免的。当我们消除了系统误差、过失误差，通过实验设计将个体差异控制在合理的范围中，则在最后的测试结果中，将只留下随机误差和条件误差。

在上述的例 6.1 中，有人可能用表 6-1 中的各点出现的频数不相等的数据来作出该骰子是不均匀的结论，这是主观的和不科学的。在掷骰子的过程中，由于受到许多不可控的偶然因素的影响，各点数出现的频数不等是显而易见的，所以，不能光凭数据的表象来做出主观的臆断。

在例 6.2 中，许多人喜欢用试验结果的数值大小来直接得出一种训练方法比另一种训练方法的效果更好的结论。这显然是不可以的。这是因为，即使两组都采用相同的训练方法，他们的均数也不一定正好完全一样。成绩的提高，不仅与训练方法有关，而且还取决于其他一些因素，如运动员的饮食、作息时间、健康状况等，在两组人数均不多的情况下，这种因素的偶合可能对某一个组比较有利。因此，问题就归结到要查明成绩提高的均值究竟是一种偶然现象，还是反映了一种训练方法比另一种训练方法更为有效。

同理，在例 6.3 中，我们也不能仅凭率的大小来得出接种疫苗后发病率明显低于不接种疫苗的对照组的结论。

而在例 6.4 中，我们已经无法根据数据的表面值的大小来直接判断它是否服从正态分布了。

哲学家卡尔·波尔说过，“无知是人类的本源”。由于人类知识的局限性，我们不可能对影响测试结果的所有因素都能了解，并能控制，因此，只要有不可控的偶然因素的存在，观察指标的测试值与其真值之间就免不了有不相等的情况出现。

因此，试验数据表面值的大小就不能作为拒绝或不拒绝统计假设的唯一依据就不难理解了。要解决这类问题，就得要借助于统计学中的假设检验的方法。

6.1.3 何谓统计检验

做出拒绝或不拒绝统计假设的统计步骤和判定规则称为统计检验。

下面, 结合实例来说明统计检验的过程。

例 6.5 某厂有一批产品共 10000 件, 按国家规定, 这些产品应达到次品率不超过 2% 的合格率后才能出厂。现检测人员从该批产品中随机抽取 50 件, 抽得次品 4 件, 问该批产品能否出厂?

【题析】 如果该批产品的次品率 P 小于等于 2%, 则该批产品显然可以出厂。但现在该批产品的次品率 P 未知, 只知抽样产品的次品率为 $4/50$, 根据样本提供的这一信息要判断该批产品能否出厂, 等价于要推断“该批产品的次品率小于等于 2%”的统计假设。

所以做统计检验的步骤如下:

(1) 根据研究问题的性质和实际问题的需要, 作出原假设 H_0 和备择假设 H_1 。一般地, 我们都期望能推翻 H_0 , 故又把假设 H_0 称为解消假设、零假设、原假设、无效假设等, 本书中将其统称为原假设。根据原假设检验的方向性, 原假设分为双侧检验和单侧检验两种。

在实际假设中, 考虑到原假设 H_0 和备择假设 H_1 在假设检验中所承担的作用是不对称的, 对原假设的处理总是偏于保守, 所以一般把所要被拒绝的假设作为原假设, 而把拒绝原假设后, 不拒绝 (或称接受) 的与原假设相反的假设作为备择假设。

在 SPSS 中, 一般都采用双侧检验。

在本例中, H_0 : 该批产品的次品率小于等于 2%, H_1 : 该批产品的次品率大于 2%。这是一个单侧检验。

(2) 以 H_0 为前提, 确定检验 H_0 的统计量及其分布。

在本例中, 在原假设成立的前提下, 由于次品发生的概率为 0.02, 在抽样 50 件的条件下, 相当于做了 50 次独立的试验, 因此, 次品发生的次数这个随机变量, 它服从二项分布。所以统计量可选用次品出现的概率, 而分布为二项分布。

(3) 计算统计量的值并根据其分布求出原假设成立的概率。

50 件产品中抽得至少 4 件次品的概率 $= 1 - P\{50 \text{ 件产品中次品少于 } 4\}$, 在 SPSS 中, 仿第 4 章中的做法, 它可利用 $1 - \text{CDF.BINOM}(3, 50, 0.04)$ 来求得, 为 0.0177581。

我们可以拒绝原假设, 也可以不拒绝原假设。但在不拒绝原假设时, 本例原假设成立的概率只有 0.0177581, 它很小。

那么, 究竟用怎样的标准来判定是拒绝还是不拒绝原假设呢?

在上面 6.1.2 中已经提到, 当我们严格控制统计误差的来源后, 最终统计结果中的误差只包含随机误差和条件误差。

根据误差分布的理论可知, 当统计结果中只包含随机误差, 或主要以随机误差为主

时, 随机误差是不改变样本的性质的, 即样本仍服从原总体的分布。而当统计结果中只包含条件误差, 或主要以条件误差为主时, 条件误差会改变样本的性质, 由于它的存在, 样本就可能不再服从原总体的分布。

一般地, 当样本与总体处于同分布状态时, 原假设成立是个大率事件, 而当样本与总体处于不同分布状态时, 原假设成立是个小率事件。

因此, 可以用原假设成立的概率的大小来区分随机误差和条件误差的作用。

所谓小率事件, 是指发生概率很小的事件。如果一个事件其发生的概率很小, 则在一次完全随机的试验中, 可以把它看成是几乎不可能发生的事件。所以, 一般情况下, 统计学上把发生的概率小于等于 0.05 的事件看成是小率事件。这就是所谓的判断拒绝或不拒绝原假设的规则即小率事件的原则。

(4) 确定拒绝原假设准备犯错误的概率, 即确定显著性水平 α 值。一般取 0.05、0.01。双侧检验将犯错误的概率均分配在分布的两侧, 而单侧检验将犯错误的概率分配在分布与原假设反向的一侧。

本例取 $\alpha = 0.05$, 即准备在拒绝原假设时犯错误的概率为 0.05。由于本例是单侧检验, 犯错误的概率全部分配在二项分布的右侧。

(5) 给出统计结论。当 $P(H_0) > 0.05$ 时, 原假设成立不是个小率事件, 因而, 没有充分的理由去拒绝原假设, 故结论为不拒绝原假设。当 $P(H_0) < 0.05$ 时, 原假设成立是个小率事件, 根据小率事件的原则, 它在一次随机试验中是不太会出现的, 故有充分的理由拒绝原假设, 而采纳它的备择假设 H_1 。

如果概率小于给定的 α 水平 (通常是 0.05), 我们可以说结果是在统计意义上显著或它们在 0.05 水平上显著或 $p < 0.05$ 时显著。

本例中, 由于 $P(H_0) = 0.0177581 < 0.05$, 所以, 拒绝该批产品的次品率小于等于 2% 的原假设, 而认为该批产品的次品率大于 2%, 所以不符合出厂的规定。

6.1.4 假设检验的种类

在 6.1.2 中, 给出的四个例子中, 例 6.1 至例 6.3 三个例子的统计假设中是含有参数的, 而例 6.4 的统计假设中不含参数。

因此, 习惯上可按下列标准把统计假设划分为两大类: 如果假设只对总体分布中的若干个参数指定取值范围, 则称这种假设为参数性假设; 否则, 称为非参数性假设。同样, 检验问题也可划分为两大类: 在已知总体分布的具体函数形式的前提下, 只是其中若干个参数未知, 则称这种检验问题为参数检验问题, 否则称为非参数检验问题。参数检验问题中的原假设和备择假设都是参数假设。而非参数检验问题的前提中并没有指定总体分布的具体函数形式, 故原假设和备择假设都可以是参数性假设也可以是非参数性假设。

所以，假设检验的种类根据检验的问题可分为参数性假设检验和非参数假设检验两类。

6.1.5 假设检验中易犯的两类错误

假设检验中用的推断原理类似数学推理中的反证法，但它同反证法有质的区别：演绎推理中用到的反证法得到的结论是正确的、确定性的，而在统计学的假设检验中得到的结论可能是不正确的。因为小概率事件在一次试验中虽然发生的概率很小，但并非完全不可能发生。因此，在假设检验中，我们不能证明一个假设是真的或是假的，只能说如果样本提供的信息足于使我们拒绝原假设时，我们就要拒绝它，否则就没有足够的证据拒绝原假设。

原假设只存在两种可能，一种是原假设为真，另一种是原假设为假。因此，当原假设为真时，但由样本判断的结果却拒绝原假设会犯错误，此类错误称为“弃真”，也称为第一类错误。它由显著性水平 α 决定。例如，如果选择 5% 的显著性水平，则犯第一类错误率为 5%。另一种情况， α 表达了当原假设真时，犯第一类错误的条件概率。

当原假设为假时，但由样本判断的结果却接受了原假设，也会犯错误，此类错误称为“取伪”，也称为第二类错误。在备择假设下，犯第二类错误的条件概率用 β 表示。

故无论是接受原假设还是拒绝原假设，都可能要犯不同的错误。对原假设判断和犯错误情况归纳成表 6-4。

表 6-4 对原假设判断情况表

	原假设为真	原假设为假
拒绝原假设	犯第一类错误 α	正确（势：概率=1- β ）
接受原假设	正确	犯第二类错误 β

犯“弃真”错误的概率 α 与“取伪”错误的概率 β 是有关联的。若 α 减小，则 β 会增大；反之，若 α 增大，则 β 会减小。在其他条件不变的情况下，使 α 、 β 同时都减小的方法是增大样本的含量 n 。

α 在统计学上也称为拒绝域的检验水平，而把 H_0 不真时，拒绝原假设（也就是，接受备择假设）的条件概率，称为势（Power），它等于 $1-\beta$ ，其意义是当备择假设真时，反映拒绝原假设的功效大小。它可以看作是研究者寻找确实存在的关系或差异的可靠性。一个最佳的检验实际上就是在寻找这样的一个拒绝域，即当备择假设真时，使拒绝原假设的功效最大，也即使犯第二类错误的概率最小。在 SPSS 中所给出的检验方法都是某种意义下的最优检验。

6.2 一元正态总体均值差异的显著性检验

在本节涉及的是一个或多个正态总体在一个尺度变量上是否有差异的显著性检验的问题。关于总体的正态性检验参见第 2 章和第 3 章中的相关内容。

本章及以后涉及到的各种检验方法同第 1 章 1.2.6 中所提到的各种实验设计方法间，不是孤立的，而是相互有关联的。在各种特定的实验设计下取得的数据资料在满足一定条件后，便由统计学家推导产生了各种统计推断对应的检验方法。

6.2.1 单样本 t 检验

对应于第 1 章 1.2.6 中的单组设计，当从正态总体中抽取的一个样本均值需要同一个已知正态总体均值 μ_0 ，当总体标准差 σ 未知时，进行均值间差异的假设检验时，可以用 SPSS 中现成的单样本 t 检验来完成。此时的原假设为

$$H_0: \mu = \mu_0 \quad (H_1: \mu \neq \mu_0)$$

由于 σ 为未知，故用 $S^2 = \frac{\sum (X - \bar{X})^2}{n-1}$ 代替 σ^2 ，当 H_0 成立时，则统计量

$$T = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{(n')} \quad (\text{其中 } n' = n - 1, \text{ 称自由度}).$$

当 $P(H_0) < 0.05$ 时，拒绝 H_0 ，否则就不拒绝 H_0 。

例 6.6 某厂用某钢生产一种钢筋的强度 X 服从正态分布，已知 $\mu_0 = 52.00 \text{ kg/mm}^2$ ，现改变炼钢的配方，利用新法炼了 10 炉钢，从由它生产的钢筋中每炉随机抽取一根，测量得到每根的强度为：53.21, 49.78, 57.32, 52.33, 50.21, 54.55, 53.78, 52.44, 55.11, 53.32，试分析用新法炼钢后钢筋的强度是否提高？

【题析】 这是一个典型的关于从正态总体中随机抽取的一个样本均值与已知总体均值之间是否有差异的显著性检验问题，所以可用 SPSS 中单样本 t 检验来处理。具体步骤如下：

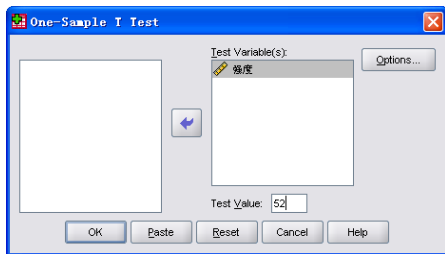


图 6-1 一个样本 t 检验对话框

(1) 在 SPSS 数据编辑窗口中，建立或打开数据文件。

由于已将本例数据建立了数据文件 data06-01.sav，所以在 SPSS 数据编辑窗口中打开 data06-01.sav。

(2) 按 Analyze → Compare Means → One-Sample T Test 顺序，打开 One-Sample T Test 对话框，见图 6-1。

(3) 选择检验变量：在左侧变量源框中选定 *强度*，并按右移箭头按钮将其移入 *Test Variable(s)*框中。

(4) 在 *Test Value*:后框中输入已知总体均值 52。

(5) 单击 *OK* 按钮，在输出窗口中输出计算结果。见表 6-5、表 6-6。

表 6-5 强度的描述统计

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
强度	10	53.2070	2.22940	.70500

表 6-6 *t* 检验结果

One-Sample Test					
	Test Value = 52				
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference
					Lower Upper
强度	1.712	9	.121	1.20700	-.3878 2.8018

(6) 结果与分析：

在表 6-5 中，显示了 *强度* 变量中一共有 10 个观察值，它的平均值为 53.2070，标准差为 2.2294，标准误为 0.705，表 6-6 中，显示了 *t* 检验的结果，从左向右依次是：*t* 值为 1.712，自由度为 9，在钢筋的强度等于 52.0 的原假设下，出现目前统计量或更极端值的双尾概率为 0.121，平均值之间的差值为 1.2070，95% 置信区间的下限为 -0.3878，上限为 2.8018。

(7) 结论：

因为 $P = 0.121 > 0.05$ ，所以没有证据拒绝原假设，即新法炼钢后钢筋的强度还是同原来一样。（还可以看 95% 的置信区间是否包含 0，如果包含 0，说明还没有理由拒绝原假设。这说明用均值差的区间估计，也是可以推断均值之间是否有统计上的显著差异的。）

例 6.7 测得某地 20 岁男性运动员和 20 岁男性大学生各 20 名的肺活量指标值见 data06-02.sav，而普查得到的正常 20 岁男子的肺活量指标的均值为 3718cm³，且 20 岁男子的肺活量指标服从正态分布，问某地 20 岁男性运动员和 20 岁男性大学生与正常 20 岁男子的肺活量之间有无统计上的显著性差异？

在 SPSS 中的解题步骤如下：

(1) 在 SPSS 数据编辑窗口中，打开数据文件 data06-02.sav。

(2) 按 *Analyze*→*Compare Means*→*One-Sample T Test* 顺序，打开 *One-Sample T Test* 对话框（图 6-1）。

(3) 选择检验变量：在左侧变量源框中同时选定 *运动员肺活量* 和 *大学生肺活量*，并按右移箭头按钮将其移入 *Test Variable(s)*框中。

(4) 在 Test Value:后框中输入已知总体均值 3718。

(5) 单击 OK 按钮, 在输出窗口中输出计算结果。见表 6-7、表 6-8。

(6) 结果与分析

在表 6-7 中, 显示了运动员肺活量和大学生肺活量变量中各有 20 个观察值, 它们的平均值分别为 4399.00 和 3667.50, 标准差分别为 531.848 和 316.758, 标准误分别为 118.925 和 70.829, 表 6-8 中, 显示了两组肺活量的 t 检验结果, 从左向右依次是: t 值分别为 5.709 和 -0.741, 自由度都为 19, 在男性运动员肺活量和男性大学生肺活量等于 3718 的原假设下, 出现目前统计量的值或更极端值的双尾概率分别为 0.000 和 0.468, 平均值之间的差值分别为 679.000 和 -52.500, 95%置信区间分别为 [430.09, 927.91] 和 [-200.75, 95.75]。

(7) 结论:

由于运动员肺活量与正常值间检验的双尾概率为 0.000, 小于 0.01, 而均值上运动员肺活量要大很多, 所以, 运动员肺活量高于正常值, 在统计上有极显著性意义, 同理, 由于大学生肺活量与正常值间检验的双尾概率为 0.468, 大于 0.05, 故大学生肺活量与正常值间无显著性差异。(还可以看 95%的置信区间是否包含 0, 如果包含 0, 说明还没有证据拒绝原假设, 如果不包含 0, 说明两者有统计上的显著性差异。由此得出的结论同 t 检验的结果是一致的。)

表 6-7 两组肺活量的描述统计

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
运动员肺活量	20	4399.00	531.848	118.925
大学生肺活量	20	3667.50	316.758	70.829

表 6-8 肺活量的 t 检验结果

One-Sample Test						
	Test Value = 3720				95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
运动员肺活量	5.709	19	.000	679.000	430.09	927.91
大学生肺活量	-.741	19	.468	-52.500	-200.75	95.75

6.2.2 独立样本 t 检验

当从两个总体中, 各随机抽取一个样本, 则这两个样本是彼此间独立没有任何关联的, 现对这两个样本, 测量同一个测试指标的值, 要看它们在这个指标上的均值之间有无显著性差异时, 这属于标准的第 1 章 1.2.6 中的成组设计, 当这两个总体分别服从各自的正态分布时 (也可由从这两个总体中随机抽取的样本服从正态分布来判定), 就可用 SPSS 中独立样本的 t 检验方法来进行推断。

由于正态总体的标准差未知, 所以, 首先要对随机抽取的样本作方差的齐次性检验, 具体方法如下:

$$H_0: \sigma_1^2 = \sigma_2^2$$

当 H_0 成立时, 则 Levene 定义统计量

$$L = \frac{(W-2) \sum W_k (\bar{Z}_k - \bar{Z})^2}{\sum_{k=1}^2 \sum_{i=1}^{n_k} W_{ki} (Z_{ki} - \bar{Z}_k)^2}$$
 服从 F 分布。

其中

$$Z_{ki} = |X_{ki} - \bar{X}_k|, \quad \bar{Z}_k = \frac{\sum_{i=1}^{n_k} W_{ki} Z_{ki}}{W_k}, \quad \bar{Z} = \frac{\sum_{k=1}^2 W_k \bar{Z}_k}{W_1 + W_2},$$

W_1 为第一个样本的容量, W_2 为第二个样本的容量, W_{ki} 为第 K 个样本中第 i 个观察值的频数。

当 $P(H_0) < 0.05$, 拒绝 H_0 , 否则不拒绝 H_0 。

例 6.8 用甲、乙两种方法冶炼某种金属, 为了检验这两种方法生产的产品中所含杂质的波动性是否有差异, 现各取一个样本, 测得数据如下:

甲	26.9	22.8	25.7	23.0	22.3	24.2	26.1	26.4	27.2	30.2	24.5	29.5	25.1
乙	22.6	22.5	20.6	23.5	24.3	21.9	20.6	23.2	23.4				

问在显著性水平 $\alpha = 0.05$ 下, 甲、乙两种方法生产的产品中所含杂质的波动性是否有显著差异?

【题析】 要检验甲、乙两种方法生产的产品中所含杂质的波动性是否有显著差异, 实际上就是要检验这两种方法所生产的产品中所含杂质的方差是否一致, 即要作方差的齐次性检验。

关于方差的齐性检验, 在第 2 章 2.4 节探索分析中已经做过介绍, 本例将在独立样本 t 检验中, 作方差的齐次性检验。具体步骤如下:

(1) 打开已经建立好的存放题中数据的数据文件 data06-03.sav。

(2) 按 Analyze→Compare Means→Independent-Samples T Test 顺序, 打开 Independent-Samples T Test 对话框, 见图 6-2。

(3) 在左侧变量源框中, 选择滚珠直径变量, 按右移箭头将其移入 Test Variable(s): 框中, 同样的方法将杂质质量变量, 移到 Grouping Variable 框中。

(4) 单击 Define Groups 按钮, 展开 Define Groups 对话框, 见图 6-3。在 Group1 中输入 1, 在 Group2 中输入 2。(或选择 Cut point, 并在其后框中输入 1 和 2 之间的任意一个值也可。) 单击 Continue 按钮返回图 6-2。

(5) 单击 OK 按钮运行, 则在输出窗口中得到想要的统计计算结果, 表 6-9。

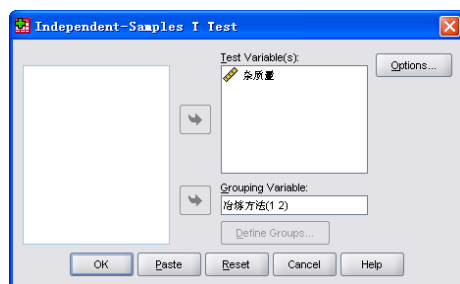
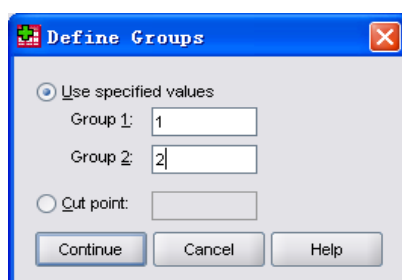
图 6-2 独立样本 t 检验对话框

图 6-3 定义组别

表 6-9 方差的齐性检验

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
				t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		F	Sig.						Lower	Upper
杂质量	Equal variances assumed	2.925	.103	3.582	20	.002	3.17350	.88584	1.32566	5.02134
	Equal variances not assumed			3.988	19.006	.001	3.17350	.79575	1.50801	4.83900

从表 6-9 的第二列 Levene 等方差检验中可知, F 值为 2.925, 在方差相等的原假设下, 出现目前统计量的值或更极端值的概率为 0.103, 因它大于 0.05, 故没有充分的证据拒绝原假设, 即认为甲、乙两种方法生产的产品中所含杂质的波动性没有显著差异。

这同用第 2 章 2.4 节探索分析中介绍的 Levene 方差齐性检验的结果是一致的。读者可自行验证。

表 6-9 中其他列的内容将在下例的输出结果分析中讨论。

当方差齐性的结果不拒绝 H_0 时, 总体均值间差别的假设检验为:

$$H_0: \mu_1 = \mu_2$$

$$t' = D/S'_D, \text{ 服从自由度为 } df = W_1 + W_2 - 2 \text{ 的 } T \text{ 分布。}$$

$$\text{其中 } D = \bar{X} - \bar{Y}, \quad \bar{X} = \sum_{i=1}^N W_i X_i / W, \quad \bar{Y} = \sum_{i=1}^N W_i Y_i / W$$

$$S'_D = S_p \sqrt{\frac{1}{W_1} + \frac{1}{W_2}}, \quad S_p^2 = \frac{(W_1 - 1)S_1^2 + (W_2 - 1)S_2^2}{W_1 + W_2 - 2}$$

当 $P(H_0) < 0.05$, 拒绝 H_0 , 否则, 没有理由拒绝 H_0 。

当方差齐性的结果否定 H_0 时, 总体均值间差别的假设检验为:

$$H_0: \mu_1 = \mu_2$$

$$t = D/S_D, \text{ 服从自由度 } df' = \frac{1}{Z_1 + Z_2} \text{ 的 } t \text{ 分布}$$

其中：
$$Z_k = \left(\frac{S_k^2 / W_k}{S_1^2 / W_1 + S_2^2 / W_2} \right)^2 / (W_k - 1), \quad S_D = \sqrt{\frac{S_1^2}{W_1} + \frac{S_2^2}{W_2}}$$
$$S_k^2 = \frac{\sum_{i=1}^{n_k} X_{ki}^2 w_{ki} - \left(\sum_{i=1}^{n_k} X_{ki} w_{ki} \right)^2 / W_k}{(W_k - 1)}$$

当 $P(H_0) < 0.05$ ，拒绝 H_0 ，否则不拒绝 H_0 。

例 6.9 从两台生产同一个型号滚珠的车床各自所生产的滚珠中，分别抽出 8 粒和 9 粒滚珠，测得其直径如下（单位：mm）：

A: 15.0, 14.5, 15.2, 15.5, 14.8, 15.1, 15.2, 14.8

B: 15.2, 15.0, 14.8, 15.2, 15.0, 14.8, 15.1, 14.9, 15.0

试问，A、B 两个车床生产的滚珠的直径有无差异？

【题析】 由于两个车床彼此独立，所以由这两台车床所生产的滚珠组成的两个样本彼此间是互相独立的。由于两样本所隶属的总体分布未作说明，故首先要作探究分析，以检查它们是否服从正态分布。

1. 建立数据文件

将上述数据在 SPSS 中建成数据文件，见 data06-04.sav。

2. 正态性检验

(1) 按 Analyze→Descriptive Statistics→Explore 顺序，打开 Explore 主对话框(图 2-99)。

(2) 在左边的变量名源框中，选中滚珠直径变量，通过中间的右移箭头将它们移到 Dependent List: 框中，用同样的方法将变量车床号，移到 Factor List 框中。在 Display 选项中，选择 Plots，单击 Plots 按钮，打开 Plots 对话框，见图 2-104。

(3) 在 Plots 对话框中，在 Boxplots 选择项中，选择 None，不输出箱图。选择 Normality Plots with tests 选项，做正态分布检验。在 Descriptive 中关闭所有选项。其他保持系统默认选项。单击 Continue 按钮，返回 Explore 主对话框。

(4) 单击 OK 按钮运行，在输出窗口中得到本例所要的正态性检验的计算结果，见表 6-10。

表 6-10 车床直径的正态性检验

Tests of Normality						
车床号	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
滚珠直径 1	.147	8	.200 [*]	.974	8	.931
2	.167	9	.200 [*]	.910	9	.319

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

从表 6-10 中可知，在车床直径服从正态分布的原假设下，出现目前统计量的值或更

极端值的概率均为 $0.200 > 0.05$ ，所以没有证据拒绝两车床生产滚珠的直径服从正态分布的原假设。

由此，要求两个车床生产的滚珠的直径有无差异，可用独立样本 t 检验来处理。

3. 独立样本 t 检验

(1) 按 Analyze → Compare Means → Independent-Samples T Test 顺序，打开 Independent-Samples T Test 对话框，见图 6-2。

(2) 在左侧变量源框中，选择滚珠直径变量，按右移箭头将其移入 Test Variable(s) 框中，同样的方法将车床号变量，移到 Grouping Variable 框中。

(3) 按 Define Groups 按钮，展开 Define Groups 对话框，见图 6-3。在 Group1: 中输入 1，在 Group2: 中输入 2。按 Continue 按钮返回图 6-2。

(4) 单击 OK 按钮运行，则在输出窗口中得到统计计算结果，见表 6-11，表 6-12。

表 6-11 分组统计量

Group Statistics				
	车床号	N	Mean	Std. Deviation
滚珠直径	1	8	15.0125	.30909
	2	9	15.0000	.15000

表 6-12 独立样本 t 检验

Independent Samples Test									
		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
滚珠直径	Equal variances assumed	3.544	.079	.108	15	.915	.01250	.11559	-.23386 .25886
	Equal variances not assumed			.104	9.860	.919	.01250	.12017	-.25578 .28078

4. 结果与分析

在表 6-11 中，显示了分组统计的结果，从左到右分别列出了 A、B 两个样本的含量为 8、9，平均值分别为 15.0125、15.0000，标准差分别为 0.30909、0.15000，标准误分别为 0.10928、0.05000。

在表 6-12 中，左边首先显示了 Levene 的方差齐性检验的 F 值为 3.544，在方差齐性的原假设下，出现目前统计量的值或更极端值的概率为 0.079，大于 0.05，所以还没有足够的证据拒绝方差齐性的假设，因此，采用等方差条件下的 t 检验结果，也即采用由矩形框框定的值，来解释 t 检验的结果。 T 值为 0.108，自由度为 15，在两个车床生产的滚珠的直径相等的原假设下，出现目前统计量的值或更极端值的概率为 0.915，平均值之间的差值为 0.01250，平均值差值的标准误为 0.11559，两均值差值的 95% 的置信区间的下限为 -0.23386，上限为 0.25886。

5. 结论

因为 t 检验中，在两个车床生产的滚珠的直径相等的原假设下，双尾概率 0.915 大于 0.05（或两均值差值的 95% 的置信区间 $[-0.23386, 0.25886]$ 包含 0），所以，现有证据不支持拒绝原假设，即 A、B 两个车床生产的滚珠的直径无统计学意义上的差异。

例 6.10 某一橡胶配方中，原用氧化锌 5 克，现减为 1 克，用这两种配方作一批试验，测得橡胶伸长率如下：

5 克	540	533	530	545	531	541	529	534	525	
1 克	565	577	580	588	556	538	560	528	570	561

问这两种配方伸长率的总体均值间有无显著性差异？

1. 建立数据文件

将例 6.10 中的数据在 SPSS 中建成数据文件，见 data06-05.sav。

2. 进行正态性检验

按例 6.9 中方法进行正态性检验，得表 6-13。从表中可知，两种配方下的橡胶伸长率服从正态分布的概率都大于等于 0.2，故不拒绝橡胶伸长率服从正态分布的假设。

表 6-13 橡胶伸长率的正态性检验结果

Tests of Normality						
配方类型	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
橡胶伸长率 氧化锌5克	.180	9	.200 [*]	.953	9	.718
氧化锌1克	.167	10	.200 [*]	.954	10	.720

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

3. 独立样本 t 检验

（1）按 Analyze → Compare Means → Independent-Samples T Test 顺序，打开 Independent-Samples T Test 对话框，见图 6-2。

（2）在左侧变量源框中，选择橡胶伸长率变量，按右移箭头将其移入 Test Variable(s) 框中，同样的方法将配方类型号变量，移到 Grouping Variable 框中。

（3）单击 Define Groups 按钮，展开 Define Groups 对话框，见图 6-3。在 Group1 中输入 1，在 Group2 中输入 2。按 Continue 按钮返回图 6-2。

（4）单击 OK 按钮运行，则在输出窗口中得到统计计算结果，见表 6-14，表 6-15。

表 6-14 分组统计量

Group Statistics				
配方类型	N	Mean	Std. Deviation	Std. Error Mean
橡胶伸长率 氧化锌5克	9	5.3422E2	6.49573	2.16524
氧化锌1克	10	5.6230E2	18.46949	5.84057

表 6-15 独立样本 t 检验

Independent Samples Test									
		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
									Lower Upper
橡胶伸长率	Equal variances assumed	4.530	.048	-4.316	17	.000	-28.07778	6.50517	-41.80250 -14.35306
	Equal variances not assumed			-4.508	11.402	.001	-28.07778	6.22900	-41.72904 -14.42652

4. 结果与分析

在表 6-14 中, 显示了分组统计的结果, 从左到右分别列出了两种配方下的两个样本的含量为 9、10, 平均值分别为 534.22、562.3, 标准差分别为 6.49573、18.46949, 标准误差分别为 2.16524、5.84057。

在表 6-15 中, 左边首先显示了 Levene 的方差齐性检验的 F 值为 4.530, 在两个样本方差齐性的原假设下, 出现目前统计量的值或更极端值的概率为 0.048, 小于 0.05, 所以有足够理由拒绝方差齐性的假设, 因此, 采用不等方差条件下的 t 检验结果, 也即采用由矩形框框定的值, 来解释 t 检验的结果。 T 值为 -4.508, 修正自由度为 11.402, 在两种配方伸长率的总体均值相等的原假设下, 出现目前统计量的值或更极端值的概率为 0.001, 平均值之间的差值为 -28.07778, 平均值差值的标准误为 6.22900, 两均值差值的 95% 的置信区间的下限为 -41.72904, 上限为 -14.42652。

5. 结论

因为 t 检验中, 双尾概率 $P=0.001$ 小于 0.05 (或两均值差值的 95% 的置信区间 $[-41.72904, -14.42652]$ 不包含 0), 所以, 两种配方伸长率的总体均值间的差异在统计上有显著性意义。

6.2.3 配对样本 t 检验

当我们用第 1 章 1.2.6 中的配对设计方法, 获取两个样本的数据资料时, 无论它是采用自身配对设计还是采用同源配对设计或条件相近配对设计, 此时应把两个总体之间看成是不独立, 如在同一个被试对象身上, 在试验前、后测得的两个样本的数据之间是不独立的, 前、后成对数据之间存在关联, 即前测数据会对后面的结果会有影响, 此时, 当两个总体分别服从正态分布时, 要检验两个不独立总体均值之间是否有差异的显著性检验就不能再用前面介绍的方法进行处理。

设 (X_i, Y_i) 分别是两个相关正态总体中抽出的两个成对出现的样本, 则我们可以用两个成对数据之间的差值 $d_i = x_i - y_i$ 构成一个新样本 D , d_i 服从正态分布。在两个正态总体均值间不存在显著性差异时, 等价于要检验新样本的均值 \bar{d} 是否服从 $\mu_0 = 0$ 的正态总体。因此, 可作如下的原假设

$H_0: \bar{d} = \mu_0 = 0$

在原假设成立的条件下，则统计量 $T = \frac{\bar{d} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{\bar{d}}{\frac{S}{\sqrt{n}}}$ 服从自由度 $n' = n - 1$ 的 t 分布。

当 $P(H_0) < 0.05$ 时，拒绝 H_0 。

例 6.11 从某校 20 岁男生中随机抽取 15 名学生，进行每天 1 小时的中长跑锻炼，测得他们参加锻炼前和锻炼一年后的晨脉数据如下：

锻炼前	70	76	72	63	63	66	68	72	65	65	75	66	76	68	70
锻炼后	52	54	60	64	50	55	62	55	51	58	56	58	64	56	58

试问锻炼前、后的晨脉间是否有显著性差异？

1. 建立数据文件

在 SPSS 中，将上述数据建成数据文件，见 data06-06.sav。

2. 进行数据资料的正态性检验

(1) 按 Analyze→Descriptive Statistics→Explore 顺序，打开 Explore 主对话框（图 2-99）。

(2) 在左边的变量名源框中，同时选中 锻炼前 和 锻炼后 变量，通过中间的右移箭头将它们移到 Dependent List: 框中。在 Display 选项中，选择 Plots，单击 Plots 按钮，打开 Plots 对话框，见图 2-104。

(3) 在 Plots 对话框中，在 Boxplots 选择项中，选择 None，不输出箱图。选择 Normality Plots with tests 选项，做正态分布检验。在 Descriptive 中关闭所有选项。其他保持系统默认选项。单击 Continue 按钮，返回 Explore 主对话框。

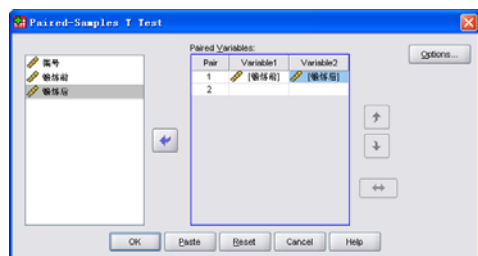
(4) 单击 OK 按钮运行，在输出窗口中得到本例所要的正态性检验的计算结果，见表 6-16。

表 6-16 锻炼前和锻炼后的晨脉的正态性检验

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
锻炼前	.150	15	.200 [*]	.927	15	.247
锻炼后	.130	15	.200 [*]	.957	15	.639

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

从表 6-16 中可知，在锻炼前和锻炼后变量服从正态分布的原假设下，出现目前统计量的值或更极端值的概率均大于 0.200 也大于 0.05，所以现有证据不支持拒绝锻炼前和锻炼后的晨脉服从正态分布的原假设。

图 6-4 配对样本 t 检验3. 进行配对样本 t 检验

(1) 按 Analyze → Compare Means → Paired-Samples T Test 顺序, 打开 Paired-Samples T Test 对话框, 见图 6-4。在变量源框中, 单击选中 锻炼前 变量, 按右移箭头, 将其移入到 Paired Variables 框中, 用同样的做法将 锻炼后 变量移入到 Paired Variables: 框中, 见图 6-4 所示。

(2) 单击 OK 按钮运行, 在输出窗中, 出现表 6-17、表 6-18、表 6-19 所示的结果。

表 6-17 配对样本统计量

Paired Samples Statistics				
	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 锻炼前	69.00	15	4.456	1.151
锻炼后	56.87	15	4.340	1.121

表 6-18 相关系数表

Paired Samples Correlations			
	N	Correlation	Sig.
Pair 1 锻炼前 & 锻炼后	15	.148	.599

表 6-19 配对样本 t 检验

Paired Samples Test									
		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
					Mean	Std. Deviation			
Pair 1	锻炼前 - 锻炼后	12.133	5.743	1.483	8.953	15.314	8.183	14	.000

4. 结果与讨论

表 6-17 显示了锻炼前和锻炼后的基本描述统计的计算结果, 从左到右依次为: 变量名(锻炼前和锻炼后)、平均值(69.00,56.87)、样本含量(15,15)、标准差(4.456,4.340)、标准误(1.151,1.121)。表 6-18 显示了锻炼前和锻炼后两个变量的相关系数及其检验, 样本量为 15, 相关系数为 0.148, 在相关系数等于 0 的原假设下, 出现目前统计量的值或更极端值的概率为 0.599。表 6-19 显示了配对样本 t 检验的结果, 从左到右依次为平均值为 12.133, 标准差为 5.743, 标准误为 1.483, 95% 的置信区间的下限为 8.953, 上限为 15.314, t 值为 8.183, 自由度为 14, 在锻炼前、后晨脉的均值相等的原假设下, 出现目前统计量的值或更极端值的概率为 0.000。

5. 结论

由于 $P=0.0000$ 小于 0.05, 故拒绝原假设, 而认为锻炼前、后的晨脉均值间差异有统计上的显著性意义, 锻炼后的晨脉明显低于锻炼前的晨脉。说明锻炼对心功能有好处。

例 6.12 测得 9 名自行车运动员在正常情况下和在附加容量为 0.5 千克筒形呼吸死区的情况下, 他们在自行车测功计上工作时的每分钟呼吸量如下:

编号	1	2	3	4	5	6	7	8	9
正常	32	31	41	41	23	37	30	22	32
附加	35	42	47	33	29	36	41	40	47

试问在这两种不同情况下他们在自行车测功计上工作时的每分钟呼吸量是否有差异？

【题析】 由于每分钟呼吸量都是从同一批被试对象身上测得，因此，两种情况下测得的数据之间是不独立的。

现将题中数据资料建成 SPSS 的数据文件，data06-07.sav。

在 SPSS 中进行检验的步骤如下：

- (1) 在 SPSS 中，打开 data06-07.sav 数据文件。
- (2) 进行数据资料的正态性检验。

方法和步骤同上例，可得表 6-20。从表 6-20 中可知，在正常和附加情况下测得的每分钟呼吸量服从正态分布的原假设下，出现目前统计量或更极端值的双尾概率为 $0.200 > 0.05$ ，所以不拒绝这两种情况下测得的每分钟呼吸量服从正态分布的假设。

表 6-20 正态检验

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
正常	.173	9	.200 [*]	.918	9	.378
附加	.129	9	.200 [*]	.954	9	.734

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

因此，本例符合配对样本 t 检验所需要的条件。

- (3) 进行配对样本 t 检验

① 按 Analyze→Compare Means→Paired-Samples T Test 顺序，打开 Paired -Samples T Test 对话框，见图 6-4。在变量源框中，单击选中 正常变量，按右移箭头，将其移入到 Paired Variables 框中，用同样的做法将 附加变量移入到 Paired Variables 框中，见图 6-4 所示。

- ② 单击 OK 按钮运行，在输出窗中，出现表 6-21、表 6-22、表 6-23 所示的结果。

- (4) 结果与讨论

表 6-21 显示了正常和附加条件情况下的基本描述统计的计算结果，从左到右依次为平均值 (32.1111,38.8889)、样本含量 (9,9)、标准差 (6.82723,6.15314)、标准误 (2.27574,2.05105)。表 6-22 显示了两个变量的相关系数及其检验，相关系数为 0.226，不相关的概率为 0.558。表 6-23 显示了配对样本 t 检验的结果，从左到右依次为平均值为 -6.77778，标准差为 8.08977，标准误为 2.69659，95%的置信区间的下限为-12.99613，上限为-0.55943， t 值为-2.513，自由度为 8，在两种条件下均值相等的原假设下，出现目前统计量的值或更极端值的概率为 0.36。

表 6-21 配对样本统计量

Paired Samples Statistics				
	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 正常	32.1111	9	6.82723	2.27574
附加	38.8889	9	6.15314	2.05105

表 6-22 相关系数表

Paired Samples Correlations			
	N	Correlation	Sig.
Pair 1 正常 & 附加	9	.226	.558

表 6-23 配对样本 t 检验

Paired Samples Test									
		Paired Differences				t	df	Sig. (2-tailed)	
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower				Upper
Pair 1	正常 - 附加	-6.77778	8.08977	2.69659	-12.99613	-5.5943	-2.513	8	.036

5. 结论

因为双尾概率 $P=0.036$ 小于 0.05 ，所以认为在这两种不同情况下他们在自行车测功计上工作时的每分钟呼吸量的均值之间的差异有统计上的显著性意义，从表 6-21 均值上可知，在附加条件下，每分钟呼吸量的均值要大于正常情况。

6.2.4 单因素方差分析

在总体服从正态分布的前提下， N 个样本对同一个测试指标上均值之间差异的显著性，似乎也可用以上介绍的独立样本均值间差异的显著性检验的方法来两两分别处理，但若用前面介绍过的两个样本间均值差异的显著性检验方法来处理本类问题的话，需要做 $N(N-1)/2$ 次两两比较，例如，有 10 个样本需要在同一个指标上做两两的 t 检验，则共需要进行 45 次检验，这一方面，显然太麻烦了，另一方面，当设定两两比较时，犯第一类错误的概率 $\alpha=0.05$ ，则 N 个独立样本两两比较时，每次比较不犯第一类错误的概率为 $0.95^{N(N-1)/2}$ ，相应犯第一类错误的概率为 $1-0.95^{N(N-1)/2}$ ，远远大于事先设定的 0.05 。因此，多个均值比较时不宜采用前面介绍的 t 检验作两两比较，应采用一种新的统计处理方法来实现。

当我们用第 1 章 1.2.6 中的单因素多水平设计方法，获取 k 个样本组成的 k 组在同一个试验指标上的数据资料后，如果 k 组样本资料能服从正态分布，则单因素方差分析是解决此类问题的关键。

6.2.4.1 单因素方差分析的基本概念

方差分析最早由英国统计学家费歇 (R.A.Fisher) 在 1923 年提出，最初用于生物学和农业试验方面，后于 1946 年由斯内德克 (G.W.Snedecor) 进一步加以完善。为纪念费歇的杰出贡献，又把它称为 F 检验。现在它已在各个领域中都得到了广泛的应用。

方差分析最初是用来检验多个独立正态总体，在方差齐性的前提下，总体均值间的

差异是否具有统计意义的一种方法。而今对多个独立正态总体在方差不齐时，也有方法对总体均值间的差异进行显著性检验。因此，只要在满足多个总体相互间的独立性、正态性的条件下，方差分析就可用来探讨多个不同实验条件或处理方法对实验结果有无影响。

6.2.4.2 单因素方差分析中偏差平方和的分解

当对一个个体进行多次重复测定时，所得到的各次测定值并不是每次都相等，而是参差不齐的，它们之间的差异一般用观（察）测值与其均值之差来表示，称偏差。如果一个测量结果同时受到多个因素影响，则每个因素都要对测定的总偏差提供相应的贡献。若以总偏差平方和来表示测定结果的偏差大小，则总偏差平方和等于各因素形成的偏差平方和的总和。这是偏差平方和的加和性原理，它是方差分析赖以建立的基础。

设 k 个独立随机变量 x_1, x_2, \dots, x_k ，分别遵从 $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2), \dots, N(\mu_k, \sigma_k^2)$ 的正态分布，现从各总体中随机抽取样本含量为 n_i 的样本，得到样本观测值 $x_{i1}, x_{i2}, \dots, x_{in_i}$ ， $i=1, 2, \dots, k$ 。见表 6-24。

表 6-24 单因素多水平实验设计后取得的数据格式

组 1（水平 1）	X_{11}	X_{1n_1}
权重 w	w_{11}	w_{1n_1}
组 2（水平 2）	X_{21}	X_{2n_2}
权重 w	w_{21}	w_{2n_2}
\vdots	\vdots	\vdots	\vdots
组 k （水平 k ）	X_{k1}	X_{kn_k}
权重 w	w_{k1}	w_{kn_k}

设 X_{lj} 表示在 l 组中第 j 个观察的值， w_{lj} 为在 l 组中第 j 个观察值的权重， W_{lj} 是 l 组中前 j 个样品的权重的和， W_l 是 l 组中所有样品的权重的和， k_l 为分组数，它用最大组别值减去最小组别值加 1 来确定，而 k' 是非空组数， n_l 为 l 组中样品的数量， W 表示所有组中样品的权重和。

则 l 组（水平 l ）对应的和及均值分别是： $T_l = \sum_{i=1}^{n_l} X_{li} w_{li}$ 和 $\bar{T}_l = \frac{T_l}{W_l}$ ，方差为：

$$S_l^2 = SS_l / (W_l - 1), \text{ 总和: } G = \sum_{i=1}^k T_i.$$

其中， $SS_l = SSQ_{l, n_l}$ ，而 SSQ_{l, n_l} 是各组递推的校正平方和，它用 Young-Cramer (1971 年)加权形式的算法来计算，即

$$SSQ_{l,i} = SSQ_{l,i-1} + \frac{w_{li} \left(X_{li} W_{l,i-1} - \sum_{j=1}^{i-1} w_{lj} X_{lj} \right)^2}{W_{l,i-1} W_{li}}$$

设总方差为 TSS ，它等于所有样本测定值与总平均值的离均差平方和，设

$$BBS = \sum_{l=1}^k T_l^2 / n_l - G^2 / W, \quad WSS = \sum_{l=1}^k SS_l$$

则可以容易地证明：

$$TSS = BBS + WSS。$$

其中 WSS 反映了同一样本内各测定值之间的变异程度，称组内偏差平方和或组内方差，它代表了试验误差（随机误差）的大小， BBS 反映了样本均值与总体均值间的变异程度，称为组间偏差平方和或组间方差，它代表了分组因素效应的大小，即不同的处理造成的差异。

在同样的波动程度下，测定数据越多，计算出的偏差平方和越大，因此，仅用偏差平方和来反映测定数据波动的大小还是不够的，还需要考虑测定数据的自由个数即自由度对偏差平方和带来的影响。

各种自由度由以下公式计算

总自由度： $W-1$

组间自由度： $k'-1$

组内自由度： $W-k'$

公式中 N 表示观测值的总数， m 表示因素所分的水平数。

由各个偏差平方和与其相应的自由度之比，可得各种因素影响的平均效应，也即总方差、组间方差、组内方差的估计值

$$\text{总方差 (TSSM)} = \frac{BBS + WSS}{W - 1}$$

$$\text{组间方差 (BBSM)} = \frac{BBS}{k' - 1}$$

$$\text{组内方差 (WSSM)} = \frac{WSS}{W - k'}$$

在 SPSS 中，当使用者用矩阵数据输入，提供了各组的权重的和 (W_l)、均值 (\bar{T}_l) 和标准差 (S_l) 时，则可用这些值计算得到

$$T_l = W_l \bar{T}_l, \quad SS_l = (W_l - 1) S_l^2 \text{ 和 } G = \sum_{i=1}^k T_i。$$

如果用户提供了合并方差 S_p^2 和它的自由度 (D) 代替单个 S_l ，则组内平方和为：

$$WSS = S_p^2 D$$

6.2.4.3 方差齐性时的单因素方差分析

从有共同方差 σ^2 的 k 个正态总体中各抽一个样本，如果 $H_0: \mu_1 = \mu_2 = \dots = \mu_k = \mu$ 成立，则 k 个正态总体既具有共同的方差 σ^2 ，又具有共同的均值，因此，从 k 个完全相同的正态总体中各抽取一个样本，就相当于从同一总体中抽取 k 个样本，因此，在原假设成立的条件下，总方差估计值、组间方差估计值、组内方差的估计值具有相同的期望值 σ^2 ，换言之， $BBSM$ 与 $WSSM$ 都是 σ^2 的无偏估计值，所以两者之比： $F = \frac{BBSM}{WSSM} \approx 1$ ，近似服从 $F(k'-1, W-k')$ 分布。当原假设成立的概率 $P(H_0) < 0.05$ ，则拒绝原假设，而接受样本来自不同的正态总体的备择假设，说明至少有两个总体的均值间存在显著性差异，也即不同水平的处理造成均值的差异是有统计意义的。否则，当 $P(H_0) > 0.05$ 时，不拒绝样本来自相同总体的原假设，即没有足够的证据说明不同水平处理间有显著性差异。

综上所述，可以得到单因素方差分析表，见表 6-25。

表 6-25 单因素方差分析表

方差来源	平方和	自由度	均方	F	临界值
组间方差 (BBS)	$\sum_{i=1}^k T_i^2 / n_i - G^2 / W$	$k' - 1$	$\frac{BBS}{k' - 1}$	$\frac{BBS}{k' - 1} / \frac{WSS}{W - k'}$	$F_{\alpha}(k' - 1, W - k')$
组内方差 (WSS)	$\sum_{i=1}^k SS_i$ ($S_p^2 D$ 由矩阵输入)	$W - k'$ (D)	$\frac{WSS}{W - k'}$		
总方差 (TSS)	$BBS + WSS$	$W - 1$			

例 6.13 为了研究三种不同的铅球教学方法的效果，将某年级三个班中，同龄的各种运动能力基本相同的男生随机分成三个组，分别按以下三种不同的方法进行教学，方法一、方法二、方法三。在三个月后，经过多次课的教学，以同样的标准测得各组成绩，见表 6-26。试问这三种方法间有无差异？

【题析】 由于参与三组的被试对象的原始条件之间是基本相同的，因此被试对象对实验结果的影响可以不用考虑，因此，本例像是一个因素（教学方法因素）3 个水平（三种不同的教法）的均值间是否存在差异的显著性检验的问题。

可否用单因素方差分析来处理，最关键的是看它们的总体是否服从正态分布。

在 SPSS 中，对类似本例的具体操作过程如下：

1. 建立数据文件

用本例数据建立的数据文件名为“data06-08.sav”。在本数据文件中有两个变量，一个是标准数值型名义测度的教学方法变量，并对它定义了值标签，1=第一种教学方法，2=第二种教学方法，3=第三种教学方法，另一个是标准数值型尺度测度的得分变量。

2. 进行样本数据资料的正态性检验和方差的齐性检验

用第2章2.4节探索分析中介绍的步骤，可得表6-27、表6-28。

从表6-27中可知，在三种教法得分分别服从正态分布的原假设下，出现目前统计量的值或更极端值的概率的概率分别为0.200、0.200、0.112都大于0.05，所以现有证据不支持拒绝三种教法得分服从正态分布的假设。

从表6-28方差齐性检验中可见，在方差齐性的原假设下，出现目前统计量的值或更极端值的概率为0.115大于0.05，所以现有证据不支持拒绝方差齐性的原假设。

表 6-26 三种教学方案下的测试结果

编号	A1	A2	A3
1	5.73	8.88	5.5
2	6.45	6.85	6.06
3	6.72	5.36	5.5
4	5.55	8.62	5.6
5	5.33	5.65	6.2
6	5.45	6.86	5.12
7	6.5	5.98	6.1
8	5.27	6.68	5.45
9	6.03	6.84	6.3
10	5.17	7.8	5.25
11	5.16	6.98	5.15
12		7.52	5.24
13		6.95	5.6
14		7.4	
15		7.2	

表 6-27 三种教法得分的正态性检验

Tests of Normality						
教学方法	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
得分 第一种教学方法	.188	11	.200 [*]	.879	11	.101
第二种教学方法	.155	15	.200 [*]	.954	15	.593
第三种教学方法	.212	13	.112	.895	13	.114

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

表 6-28 方差的齐性检验

Test of Homogeneity of Variance					
		Levene Statistic	df1	df2	Sig.
得分	Based on Mean	2.298	2	36	.115
	Based on Median	1.943	2	36	.158
	Based on Median and with adjusted df	1.943	2	24.229	.165
	Based on trimmed mean	2.284	2	36	.116

3. 进行方差分析

按 Analyze→Compare Means→One-Way ANOVA 顺序,打开 One-Way ANOVA 对话框,见图 6-5。在变量源框中,单击选中得分变量,按右移箭头按钮,将它移入到 Dependent List 下框中,用同样的方法将教学方法变量移入到 Factor 下框中。

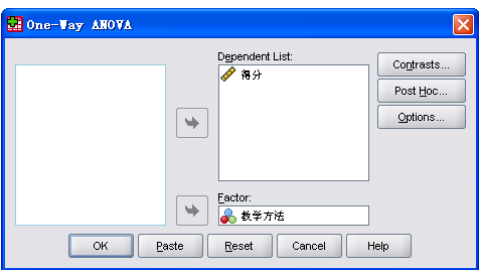


图 6-5 单因素方差分析对话框

在方差齐性的前提下,只要保持系统默认选项,直接单击 OK 按钮运行,则在输出窗口中出现方差齐性时的单因素方差分析的计算结果,见表 6-29 和表 6-30。

表 6-29 基本描述统计

Descriptives								
得分					95% Confidence Interval for Mean			
					Lower Bound	Upper Bound		
	N	Mean	Std. Deviation	Std. Error			Minimum	Maximum
第一种教学方法	11	5.7600	.57379	.17301	5.3745	6.1455	5.16	6.72
第二种教学方法	15	7.0380	.96262	.24855	6.5049	7.5711	5.36	8.88
第三种教学方法	13	5.6208	.41165	.11417	5.3720	5.8695	5.12	6.30
Total	39	6.2051	.96413	.15438	5.8926	6.5177	5.12	8.88

表 6-30 单因素方差分析表

ANOVA					
得分	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	17.024	2	8.512	16.746	.000
Within Groups	18.299	36	.508		
Total	35.323	38			

4. 分析

表 6-29 列出了三种教学方法对应的样本含量、平均值、标准差、标准误、均值 95% 的置信区间的下限、上限,以及三种教学方法对应的三个样本中各自的最小值、最大值。

表 6-30 列出了单因素方差分析的整个计算过程。偏差平方和的来源分为组间、组内两部分,从表中第二列可见,组间偏差平方和为 17.024,组内偏差平方和 18.299,因此,总偏差平方和为等于组间偏差平方和+组内偏差平方和=17.024+18.299=35.323。组间自由度为 2,组内自由度为 36,总自由度为两者之和,即 36+2=38,组间方差=组间偏差平方和/组间自由度=17.024/2=8.512 (见第四列),组内方差=组内偏差平方和/组内自由度=18.299/36=0.508, F 值=组间方差/组内方差=8.512/0.508=16.746 (见第五列),在三组均

值间相等的原假设下, 出现目前统计量的值或更极端值的概率为 0.000 (见第六列)。

因此, 方差分析的结果拒绝原假设, 而认为至少有两个均值间存在着显著性差异。

6.2.4.4 方差不齐性时的单因素方差分析

方差不齐时, 用 Brown-Forsythe 检查法比上面的方差分析更为稳健。Brown-Forsythe 检查时的原假设为:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_K$$

在原假设成立时, Brown 和 Forsythe 在 1974 年提出如下的在方差不齐时检验均值相等的统计量

$$F_{BF} = \frac{\sum_{I=1}^K W_I (\bar{T}_I - \bar{G})^2}{\sum_{I=1}^K (1 - W_I/W) S_I^2} \sim F(K-1, f)$$

$$\text{其中, } \frac{1}{f} = \sum_{I=1}^K c_I^2 / (W_I - 1), \quad c_I = \frac{(1 - W_I/W) S_I^2}{\sum_{I=1}^K (1 - W_I/W) S_I^2}$$

当我们观察 F_{BF} 的分母时, 我们可以看到它用 $S_{pool}^2 = \sum_{I=1}^K \omega_I^* S_I^2$ 估计联合方差, 其中,

$$\omega_I^* = \frac{(W - W_I)}{W(K-1)}$$

需要注意的是, 当所有组标准差为 0 或任何组的样本含量小于等于 1 时, Brown-Forsythe 统计量不能计算。在一些组标准差为 0 的情况中, 可以计算统计量, 但不能用近似分布。

所以, 当 $P(H_0) < 0.05$, 拒绝原假设。

除这种方法以外, 还可以考虑使用 Welch 检验法。

在不考虑方差齐性假设的前提下, Welch 在 1947 年至 1951 年期间, 提出了均值相等的近似检验。

$$\text{其统计量: } F_{Welch} = \frac{\sum_{I=1}^k \omega_I \left[(\bar{T} - \tilde{X})^2 / (k-1) \right]}{1 + \frac{2(k-2)}{(k^2-1)} \sum_{I=1}^k \left[\left(1 - \frac{\omega_I}{\mu} \right)^2 / (W_I - 1) \right]} \text{ 近似服从 } F(k-1, f)。$$

$$\text{其中, } \omega_I = W_I / S_I^2, \quad \mu = \sum_{I=1}^k \omega_I, \quad \tilde{X} = \sum_{I=1}^k \omega_I \bar{T}_I / \mu$$

而 $f = \left[\frac{3}{k^2 - 1} \sum_{l=1}^k \left(1 - \frac{\omega_l}{\mu} \right)^2 / (W_l - 1) \right]^{-1}$

值得一提的是，由于使用在 Welch 统计量中的权重是 $\omega_l = W_l / S_l^2$ ，所以当任何一个组的标准差为 0 时不能计算统计量。此外，所有组的样本含量必须大于等于 0。

例 6.14 为反映患者免疫状态紊乱而导致造血功能障碍的程度，研究人员从轻度、重度再障贫血患者和正常人的群体中，分别随机抽取 10 人，测试其血清中可溶性 CD₂ 抗原水平（U/ml），得如下结果，见表 6-31。

表 6-31 三组可溶性 CD₂ 抗原水平（U/ml）测试结果

正常组	234	318	402	382	421	408	343	241	292	398
轻度组	509	518	555	758	845	712	585	448	753	896
重度组	851	562	918	631	653	843	659	849	762	901

试问，这三组的 CD₂ 抗原水平有无差异？

【题析】 本例中的血清中可溶性 CD₂ 抗原水平为试验指标，贫血程度为试验因素，而正常、轻度和重度是该因素所分的三个水平，所以，它是一个一元变量多水平的单因素分析的模型，可否用单因素方差分析来处理，关键看其分布情况。

在 SPSS 中，具体的操作步骤如下：

1. 在 SPSS 数据编辑窗口中，将表 6-31 中的数据建成数据文件 data06-09.sav。
2. 进行样本数据资料的正态性检验

具体做法同例 6.13。用第 2 章 2.4 节探索分析中介绍的步骤，可得表 6-32。

表 6-32 样本数据正态性检验结果

Tests of Normality

组别	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
抗原水平 正常组	.207	10	.200 [*]	.887	10	.157
轻度组	.181	10	.200 [*]	.933	10	.483
重度组	.236	10	.122	.902	10	.230

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

从表 6-32 可见，在三个样本数据都服从正态分布的原假设下，出现目前统计量的值或更极端值的概率都大于 0.05，所以现有证据不支持拒绝原假设。

因此，本例可用单因素分析的模型来分析。

3. 方差的齐性检验

虽然，用上例中用到的第三章中介绍的探索研究的方法是完全可以进行方差的齐性检验的，但为了说明进行这种检验的场合不止一处，本例将另辟新径进行多样本方差的齐性检验。

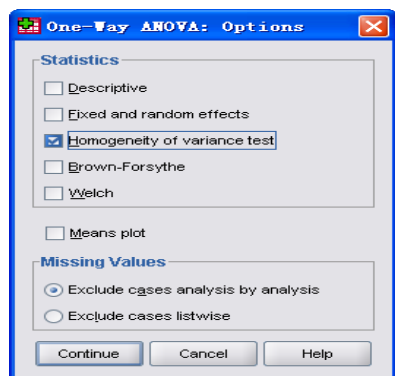


图 6-6 Options 对话框

具体步骤如下:

(1) 按 Analyze → Compare Means → One-Way ANOVA 顺序, 打开 One-Way ANOVA 对话框, 见图 6-5。在变量源框中, 单击选中抗原水平变量, 按右移箭头按钮, 将它移入到 Dependent List 下框中, 用同样的方法将组别变量移入到 Factor 下框中。

(2) 单击 Options 按钮, 打开 Options 对话框, 见图 6-6。在 Statistics 选择项中, 选择 Homogeneity of variance test 选项, 要求做多样本方差的齐性检验。其他采用系统默认值。单击 Continue 按钮, 返回 One-Way ANOVA 对话框。

(3) 按 OK 按钮执行, 在输出窗口中, 输出两张表, 第一张为方差齐性表, 见表 6-29, 第二张为方差分析表。第二张表暂时还不能分析, 故先不列出。

(4) 结果与讨论

从表 6-33 可见, Levene Statistic 的值为 5.982, 第一自由度为 2, 第二自由度为 27, 在方差齐性的原假设下, 出现目前统计量的值或更极端值的概率为 0.007。

(5) 方差齐性检验结论

因为 $P=0.007$ 小于 0.05, 所以拒绝原假设, 而认为三个样本所隶属的正态总体间的方差不齐。因此, 刚才输出结果中的方差分析表不能用。

4. 方差不齐时的单因素方差分析

(1) 按 Analyze → Compare Means → One-Way ANOVA 顺序, 打开 One-Way ANOVA 对话框, 见图 6-5。在变量源框中, 单击选中抗原水平变量, 按右移箭头按钮, 将它移入到 Dependent List 下框中, 用同样的方法将组别变量移入到 Factor 下框中。

(2) 单击 Options 按钮, 打开 Options 对话框 (见图 6-6), 在 Statistics 中, 选择 Brown-Forsythe 选项和 Welch 选项, 要求做方差不齐时的单因素方差分析。其他采用系统默认值。单击 Continue 按钮, 返回 One-Way ANOVA 对话框。

(3) 单击 OK 按钮执行, 在输出窗口中, 输出两张表, 第一张为方差分析表, 见表 6-34, 第二张为 Brown-Forsythe 检验表, 见表 6-35。

(4) 结果与讨论

在方差不齐的情形下, 用表 6-34 来解释结果显然是不合适的。应使用表 6-35 来解释。对照表 6-30 和表 6-31 不难发现 Brown-Forsythe 法与常规的方差分析法的 F 值没有区别, 都是 31.720, 第一自由度上没有区别, 都是 2, 但而第二自由度上, Brown-Forsythe 检验

表 6-33 方差齐性检验表

Test of Homogeneity of Variances

抗原水平			
Levene Statistic	df1	df2	Sig.
5.982	2	27	.007

表 6-34 方差分析表

ANOVA					
抗原水平	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	950606.667	2	475303.333	31.720	.000
Within Groups	404582.700	27	14984.544		
Total	1355189.367	29			

表 6-35 Brown-Forsythe、Welch 均值相等检验表

Robust Tests of Equality of Means				
抗原水平	Statistic ^a	df1	df2	Sig.
Welch	48.249	2	16.002	.000
Brown-Forsythe	31.720	2	21.214	.000

a. Asymptotically F distributed.

中是 21.214 小于单因素方差分析中的 27, 说明由于受到方差不齐的影响, Brown-Forsythe 检验对原检验方法进行了修正, 从而使第二自由度变小, 拒绝原假设更困难。Welch 均值相等检验则是不同于上述两者的另一种方法, 计算得到的 F 值大于前述的两者, 但修正的第二自由度又小于前述的两者, 所以有时可能会同 Brown-Forsythe 检验得出不同的结论。在知道方差不齐时, 建议使用 Brown-Forsythe 检验, 而当不知道方差齐还是不齐时, 可以使用 Welch 均值相等检验法。

(5) 结论

因为 Brown-Forsythe 检验中, P 为 0.000, 小于 0.05, 所以拒绝均值相等的原假设, 而认为至少有两个均值间存在显著性差异。本例中同 Welch 均值相等检验法的结论不矛盾。

6.2.4.5 单因素方差分析有显著性意义时两两均值间差异的多重比较

当方差分析的结果拒绝原假设, 即至少有两组的均数间存在显著性差异时, 需要对哪些均值间有显著性差异作多重比较。

多重比较的基本思想是在两组均值无显著性差异的原假设下, 计算所要比较的两组均值差在给定显著性水平 α 时的最小极差, 如果所要比较的两组均值差大于其两组均值差的最小极差, 即 $|T(q) - T(i)| \geq R_{q-i+1} M_{q,i}$, 则 $P(H_0) < \alpha$ 。

式中的 R_r 统称为极差 (Range), $M_{i,j}$ 称为极差系数。

SPSS 中几种常见多重比较法极差的计算方法, 见表 6-36。

表 6-36 常见多重比较法极差 (Range) 的产生

比较法名称	极差计算	下标范围	极差备注	α 取值
SNK	$R_r = S_{r,f}$	$r = 2, 3, \dots, k'$	$S_{r,f}$, 学生氏极差, 其中 r 是均值之间的间隔数, f 是组内均方的自由度	0.05
TUKEY	$R_r = S_{k',f}$			
TUKEYB	$R_r = \frac{S_{r,f} + S_{k',f}}{2}$			
DUNCAN	$R_r = D_{r,f}$	$r = 2, 3, \dots, k'$	极差 ($D_{r,f}$) 使用 Gebhardt (1966 年) 算法产生	0.01、0.05 和 0.10
SCHEFFE	$R_r = \sqrt{2(k'-1)F_{1-\alpha}(k'-1, f)}$			任意一个小于等于 0.05 的 α
LSD	$R_r = \sqrt{2F_{1-\alpha}(1, f)}$			
MODLSD	$R_r = \sqrt{2F_{1-\alpha'}(1, f)}$	$\alpha' = 2\alpha k'(k'-1)$		

极差系数在默认状态下用 $M_{i,j} = S_p \sqrt{\frac{1}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$ 计算, 在计算各组样本含量的调和均

值的情况下, 也可用 $M_{i,j} = S_p \sqrt{\frac{\sum_{l=1}^k \frac{1}{n_l}}{k}}$ 计算。

例 6.15 仍以例 6.13 为例, 在已知方差齐性, 单因素方差分析有显著性意义的前提下, 试分析三种教法间, 哪两种教法间有显著性差异?

在 SPSS 中, 进行多重比较的步骤如下:

1. 打开数据文件 data06-08.sav。

2. 按 Analyze→Compare Means→One-Way ANOVA 顺序, 打开 One-Way ANOVA 对话框, 见图 6-5。在变量源框中, 单击选中得分变量, 按右移箭头按钮, 将它移入到 Dependent List 下框中, 用同样的方法将教学方法变量移入到 Factor 下框中。

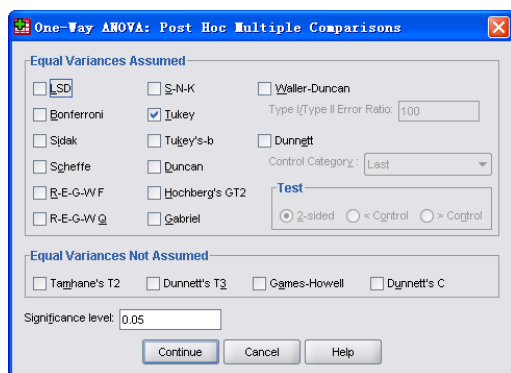


图 6-7 多重比较对话框

3. 在 One-Way ANOVA 对话框(图 6-5)中, 单击 Post Hoc 按钮, 弹出 Post Hoc Multiple Comparisons (均值差异的多重比较) 对话框, 见图 6-7。

Post hoc (后调和均值) 极差检验 (range tests) 和成对多重比较可以确定均值间的差异。极差检验识别相互间没有差异的均值的同类子集, 成对多重比较检验各个成对均值间差异, 并产生用星号象征组间均值在 0.05 水平上有显著差异的矩阵。

因为本例方差为齐性, 所以在多重比较的对话框中, 选择 Equal Variances Assumed (方差齐性) 下 14 种判断两两均值间是否有显著性差异的检验方法中的一种方法即可。

在这些方法中, 属于极差检验的方法是: Tukey's b, S-N-K (Student-Newman-Keuls)、Duncan、R-E-G-W F (Ryan-Einot-Gabriel-Welsch F test)、R-E-G-W Q (Ryan-Einot-Gabriel-Welsch test) 和 Waller-Duncan。

属于多重比较的方法是: Bonferroni、Tukey's honestly significant difference test、Sidak、Gabriel、Hochberg、Dunnnett、Scheffé 和 LSD (least significant difference)。

既属于多重标比较又属于极差检验的方法是: Tukey's honestly significant difference test、Hochberg's GT2、Gabriel 和 Scheffé。

其中 S-N-K 法是两两比较中用的最多的, 它是极差检验法, 它控制犯第一类错误的

概率为 α 。而 Tukey's honestly significant difference test 既做极差检验同时还做多重比较，而且它控制的犯第一类错误的概率比 S-N-K 法要小。

因此，本例选用 Tukey 法。其他保持系统默认选择项，即双侧检验、显著性水平 $\alpha = 0.05$ 。

单击 Continue 按钮，返回图 6-5 单因素方差分析对话框。

4. 单击 OK 按钮执行，在输出窗口中得到计算结果，见表 6-37、表 6-38。

表 6-37 多重比较

表 6-38 同类子集

Multiple Comparisons

		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
(I) 教学方法	(J) 教学方法				Lower Bound	Upper Bound
第一种教学方法	第二种教学方法	-1.27800*	.28301	.000	-1.9698	-.5862
	第三种教学方法	.13923	.29208	.883	-.5747	.8532
第二种教学方法	第一种教学方法	1.27800*	.28301	.000	.5862	1.9698
	第三种教学方法	1.41723*	.27016	.000	.7589	2.0776
第三种教学方法	第一种教学方法	-.13923	.29208	.883	-.8532	.5747
	第二种教学方法	-1.41723*	.27016	.000	-2.0776	-.7589

Tukey HSD

教学方法	N	Subset for alpha = 0.05	
		1	2
第三种教学方法	13	5.6208	
第一种教学方法	11	5.7600	
第二种教学方法	15		7.0380
Sig.		.875	1.000

Means for groups in homogeneous subsets are displayed.

*. The mean difference is significant at the 0.05 level.

5. 结果与分析

在表 6-37 中，第一列列出的是均值两两比较的组成，第二列列出了两个均值之间的差值，第三列列出了两个均值之间差值的标准误，第四列列出了在两个均值相等的原假设下，出现目前统计量的值或更极端值的概率。第五列列出了两个均值之间差值的 95% 的置信区间的下限和上限的值。两均值差值上有*号标注的，说明两个均值间的差异有统计上的显著性意义。

在表 6-38 中，列出了同类子集的分布情况。

由此可见，第二种教法的均值同第一、第三种教法的均值间的差异有统计上的显著性意义，而第一、第三种均值间差异没有显著性意义。

6. 结论

多重比较结果表明，第二种教学方法同第一、第三种教学方法的差异有统计上的显著性意义，又第二种教学方法的均值最大，故可认为第二种教法最有效。

例 6.16 仍以例 6.14 为例，在已知方差不齐，单因素方差分析有显著性意义的前提下，试分析三组间，哪两组在血清中可溶性 CD₂ 抗原水平指标上有显著性差异？

具体操作步骤如下：

1. 打开数据文件 data06-09.sav。

2. 按 Analyze→Compare Means→One-Way ANOVA 顺序，打开 One-Way ANOVA 对话框，见图 6-5。在变量源框中，单击选中抗原水平变量，按右移箭头按钮，将它移入到 Dependent List 下框中，用同样的方法将组别变量移入到 Factor 下框中。

3. 单击 Post Hoc 按钮, 弹出 Post Hoc Multiple Comparisons (均值差异的多重比较) 对话框 (见图 6-7)。

本例已经得到的前提条件是接受方差不齐的假设, 因此, 要进行均值差异的多重比较, 必须在图 6-7 中选择 Equal Variances Not Assumed(方差不齐)选项。

在方差不齐的条件下, 共有 Tamhane's T2、Dunnett's T3、Games-Howell 和 Dunnett's C 4 种判断两两均值间是否有显著性差异的检验方法。

一般认为 Games-Howell 法稍好一些, 故选择 Games-Howell, 单击 Continue 按钮返回 One-Way ANOVA 对话框。

4. 单击 OK 按钮执行, 在输出窗口中, 输出两张表, 第一张为方差分析表, 见表 6-34, 第二张为方差不齐时的多重比较, 见表 6-39。

表 6-39 方差不齐时的多重比较

		Multiple Comparisons				
		Games-Howell				
		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
(I) 组别	(J) 组别				Lower Bound	Upper Bound
正常组	轻度组	-314.00000 [*]	53.66892	.000	-456.3868	-171.6132
	重度组	-419.00000 [*]	45.84876	.000	-539.0224	-298.9776
轻度组	正常组	314.00000 [*]	53.66892	.000	171.6132	456.3868
	重度组	-105.00000	63.31086	.249	-267.1142	57.1142
重度组	正常组	419.00000 [*]	45.84876	.000	298.9776	539.0224
	轻度组	105.00000	63.31086	.249	-57.1142	267.1142

*. The mean difference is significant at the 0.05 level.

5. 结果与讨论

关于表中各列所列表头内容, 同例 6.15 中表 6-37 的解释。根据表中两两比较时均值差值上的*号或其后的在均值相等的原假设下, 出现目前统计量的值或更极端值的概率 (Sig.) 的值小于 0.05 可以看出, 正常组与轻度组和重度组之间差异都有统计上的显著性意义。而重度组与轻度组之间差异没有统计上的显著性意义。

6. 结论

重度组和轻度组与正常组之间在均值上的差异有统计上极显著性意义。

6.2.4.6 单因素方差分析有显著性意义时多组均值间的对照比较

1. 多项式比较 (Polynomial Contrasts)

Speed 于 1976 年提出了单因素方差分析中的多项式比较方法, 他将组间平方和分解为线性、二次、三次或更高次的多项式, 这样在方差分析结果中不仅可以输出组间偏差平方和, 还可以显示组间偏差平方和的各个分解结果以及 F 统计量和相伴概率。它被称

为单因素方差分析的多项式检验。

(1) 未加权对照和统计量

对第 q 次对照, F 统计量用下式计算

$$F = \frac{\left[\sum_{i=1}^k (\bar{T}_i - \bar{G}) c_{i,q} \right]^2}{\sum_{i=1}^k c_{i,q}^2 / W_i} / WSSM$$

其中 $WSSM$ 是组内均方。显著性水平从自由度 1 和 $W - k'$ 的 F 分布中获得。

上式中, 正交多项式的系数用以下的关系递推计算的:

$$c_{i,q} = (i - A_q) c_{i,q-1} - C_q c_{i,q-2}, \quad i=1, 2, \dots, k, \quad q=1, 2, \dots, NP。$$

$$A_q = \frac{\sum_{i=1}^k i c_{i,q-1}^2}{\sum_{i=1}^k c_{i,q-1}^2}, \quad \text{对于 } q \geq 2, \quad C_q = \frac{\sum_{i=1}^k c_{i,q-1}^2}{\sum_{i=1}^k c_{i,q-2}^2}, \quad \text{对于 } q=1, \quad C_q = 0。$$

(2) 加权对照和统计量

对于第 q 次正交多项成分贡献的检验是根据: $F = D_q / WSSM$

$$\text{其中, } D_q = \frac{\left(\sum_{i=1}^k w_i \bar{T}_i d_{i,q} \right)^2}{\sum_{i=1}^k w_i d_{i,q}^2}$$

显著性水平的计算是从自由度 1 和 $W - k'$ 的 F 分布中获得。

对于同第 q 次多项式偏差的检验是根据: $F = DD_q / WSSM$

$$\text{其中, } DD_q = \left(BSS - \sum_{j=1}^q D_j \right) / (k' - q - 1)$$

显著性水平用自由度为 $k' - q - 1$ 和 $W - k'$ 的 F 分布计算。打印的最高次数是 $k' - 2$ 和 5 中的最小值。

上面的公式中, 对第 q 次多项成分的对照, 用以下递推关系计算:

$$d_{i,q} = (i - A'_q) d_{i,q-1} - C'_q d_{i,q-2}, \quad i=1, 2, \dots, k, \quad q=1, 2, \dots, NP$$

$$A'_q = \frac{\sum_{i=1}^k i W_i d_{i,q-1}^2}{\sum_{i=1}^k W_i d_{i,q-1}^2}, \quad \text{对于 } q \geq 2, \quad C'_q = \frac{\sum_{i=1}^k i W_i d_{i,q-1} d_{i,q-2}}{\sum_{i=1}^k W_i d_{i,q-2}^2}, \quad \text{对于 } q=1, \quad C'_q = 0$$

2. 实例分析

例 6.17 接例 6.16 在重度组和轻度组与正常组之间在均值上有极显著性差异的前提下,可否认为重度再障贫血患者的血清中可溶性 CD₂ 抗原水平的均值是正常人的 2.4 倍,而轻度再障贫血患者的血清中可溶性 CD₂ 抗原水平的均值是正常人的 2 倍?

在 SPSS 中操作步骤如下:

① 打开数据文件 data06-09.sav。

② 按 Analyze→Compare Means→One-Way ANOVA 顺序,打开 One-Way ANOVA 对话框,见图 6-5。在变量源框中,单击选中抗原水平变量,单击右移箭头按钮,将它移入到 Dependent List 下框中,用同样的方法将组别变量移入到 Factor 下框中。

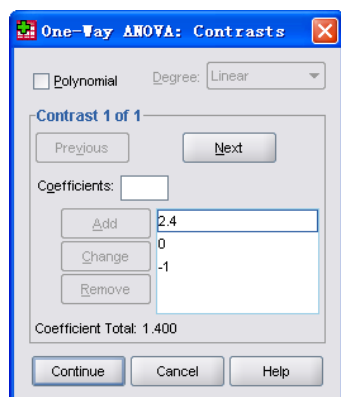


图 6-8 Contrasts 对话框

③ 单击 Contrasts 按钮,打开 Contrasts 对话框,见图 6-8。

④ 输入均值比较时的系数

单击 Coefficients 后框,输入 2.4,单击 Add 按钮,将 2.4 移入系数框,在 Coefficients 后框输入 0,单击 Add 按钮,将 0 移入系数框,再在 Coefficients 后框输入-1,单击 Add 按钮,将-1 移入系数框,完成正常组与重度组均值比较时的系数设置。

单击 Next 按钮,进行正常组与轻度组均值比较的系数设置,同重度组的操作过程一样,依次输入 2、-1、0。

单击 Continue 按钮,返回 One-Way ANOVA 对话框。

⑤ 单击 OK 按钮执行,则在输出窗口中得到输出结果,共三张表,第一张是方差分析表,见表 6-37,第二张是比较的系数表,见表 6-40,第三张是比较检验表,见表 6-41。

3. 结果与讨论

表 6-40 列出了比较时的两两系数关系,即所作的原假设分别为: $2.4 \times \text{正常组均值} - \text{重度组均值} = 0$ 和 $2 \times \text{正常组均值} - \text{轻度组均值} = 0$ 。表 6-41 给出了在方差齐性和方差不齐时的两种情况下的两两比较结果,本例要用方差不齐,即最后两行的结果,两组比较时观测的显著性水平都大于 0.05。

表 6-40 比较检验表

Contrast Coefficients			
Contrast	组别		
	正常组	轻度组	重度组
1	2.4	0	-1
2	2	-1	0

表 6-41 比较检验表

Contrast Tests						
		Contrast	Value of Contrast	Std. Error	t	Sig. (2-tailed)
抗原水平	Assume equal variances	1	62.4600 ^a	1.0064E2	.621	.540
		2	29.9000 ^a	86.55791	.345	.732
	Does not assume equal variances	1	62.4600 ^a	66.48834	.939	.361
		2	29.9000 ^a	65.89120	.454	.655

a. The sum of the contrast coefficients is not zero.

4. 结论

没有充分的理由可以拒绝原假设，因而可以认为重度再障贫血患者的血清中可溶性 CD₂ 抗原水平的均值是正常人的 2.4 倍，而轻度再障贫血患者的血清中可溶性 CD₂ 抗原水平的均值是正常人的 2 倍的假设。

6.2.4.7 单因素方差分析综合实例分析

例 6.18 为了探讨不同缺氧方式影响肺泡表面活性物质代谢的规律，用家兔作为被试对象，实验因素为 A（处理组别），分别为 A₁（对照组）、A₂（急性缺氧组）、A₃（间断缺氧 5d 组）、A₄（间断缺氧 15d 组），并将 36 只健康成年家兔随机地分到 4 个处理组中，每组 9 只家兔，试验指标为肺泡支气管灌洗液中 5 种磷脂的相对含量，即溶血磷脂酰碱（LPC）、磷脂酰碱（PC）、磷脂酰甘油（PG）、神经鞘磷脂（SPH）和磷脂酰乙醇胺（PE）。实验结果存放在 data06-10.sav 中，试问不同实验条件对这 5 个指标有无影响？

在 SPSS 中，具体的操作步骤如下：

- 1. 在 SPSS 数据编辑窗口中，打开数据文件 data06-10.sav。
- 2. 进行样本数据资料的正态性检验。

具体做法同例 6.13。用第 2 章 2.4 节探索分析中介绍的步骤，可得表 6-42。

表 6-42 样本数据正态性检验结果

		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
溶血磷脂酰碱	对照组	.145	9	.200 [*]	.980	9	.966
	急性缺氧组	.137	9	.200 [*]	.963	9	.832
	间断缺氧 5d 组	.157	9	.200 [*]	.953	9	.727
	间断缺氧 15d 组	.174	9	.200 [*]	.968	9	.876
磷脂酰碱	对照组	.210	9	.200 [*]	.895	9	.227
	急性缺氧组	.134	9	.200 [*]	.952	9	.708
	间断缺氧 5d 组	.246	9	.123	.873	9	.132
	间断缺氧 15d 组	.166	9	.200 [*]	.935	9	.526
磷脂酰甘油	对照组	.178	9	.200 [*]	.907	9	.296
	急性缺氧组	.176	9	.200 [*]	.935	9	.535
	间断缺氧 5d 组	.164	9	.200 [*]	.937	9	.554
	间断缺氧 15d 组	.132	9	.200 [*]	.963	9	.827
神经鞘磷脂	对照组	.155	9	.200 [*]	.957	9	.770
	急性缺氧组	.226	9	.200 [*]	.874	9	.135
	间断缺氧 5d 组	.193	9	.200 [*]	.887	9	.186
	间断缺氧 15d 组	.200	9	.200 [*]	.910	9	.318
磷脂酰乙醇胺	对照组	.115	9	.200 [*]	.976	9	.941
	急性缺氧组	.215	9	.200 [*]	.928	9	.466
	间断缺氧 5d 组	.252	9	.103	.917	9	.370
	间断缺氧 15d 组	.123	9	.200 [*]	.965	9	.850

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

从表 6-42 可见，四个样本 5 个指标的数据，在服从正态分布的原假设下，出现目前统计量的值或更极端值的概率（Sig.）的值都大于 0.05，所以没有足够的理由拒绝原假设。因此，本例可用单因素分析的模型来分析。

3. 方差的齐性检验

（1）按 Analyze→Compare Means→One-Way ANOVA 顺序，打开 One-Way ANOVA 对话框，见图 6-5。在变量源框中，选中溶血磷脂酰碱、磷脂酰碱、磷脂酰甘油、神经鞘

磷脂和磷脂酰乙醇胺变量,按右移箭头按钮,将它移入到 Dependent List 下框中,用同样的方法将实验因素变量移入到 Factor 下框中。

(2) 单击 Options 按钮,打开 Options 对话框,见图 6-6。在 Statistics 选择项中,选择 Homogeneity of variance test 选项,要求做多样本方差的齐性检验。其他采用系统默认值。单击 Continue 按钮,返回 One-Way ANOVA 对话框。

(3) 单击 OK 按钮执行,在输出窗口中,输出两张表,第一张为方差齐性表,见表 6-43,第二张为方差分析表。第二张表暂时还不能分析,故先不列出。

表 6-43 方差齐性检验表

	Levene Statistic	df1	df2	Sig.
溶血磷脂酰碱	7.932	3	32	.000
磷脂酰碱	3.551	3	32	.025
磷脂酰甘油	6.033	3	32	.002
神经鞘磷脂	6.760	3	32	.001
磷脂酰乙醇胺	9.958	3	32	.000

(4) 结果与讨论

从表 6-43 可见,5 个指标方差齐性的 P 值均小于 0.05。

(5) 方差齐性检验结论

因为 5 个指标在方差齐性的原假设下,出现目前统计量的值或更极端值的概率 (Sig.) 的值均小于 0.05,所以拒绝原假设,而认为 4 个样本 5 个指标所隶属的正态总体间的方差不齐。因此,刚才输出结果中的方差分析表不能用。

4. 方差不齐时的单因素方差分析

(1) 按 Analyze→Compare Means→One-Way ANOVA 顺序,打开 One-Way ANOVA 对话框,见图 6-5。

(2) 单击 Options 按钮,打开 Options 对话框(见图 6-6),在 Statistics 中,选择 Brown-Forsythe 选项和 Welch 选项,要求做方差不齐时的单因素方差分析。其他采用系统默认值。单击 Continue 按钮,返回 One-Way ANOVA 对话框。

(3) 单击 OK 按钮执行,在输出窗口中,输出两张表,第一张为方差分析表,见表 6-44,第二张为 Brown-Forsythe 检验表,见表 6-45。

表 6-44 方差分析表

		Sum of Squares	df	Mean Square	F	Sig.
溶血磷脂酰碱	Between Groups	72.618	3	24.206	16.379	.000
	Within Groups	47.293	32	1.478		
	Total	119.911	35			
磷脂酰碱	Between Groups	853.018	3	284.339	2.539	.074
	Within Groups	3583.386	32	111.981		
	Total	4436.404	35			
磷脂酰甘油	Between Groups	5307.634	3	1769.211	31.308	.000
	Within Groups	1808.339	32	56.511		
	Total	7115.973	35			
神经鞘磷脂	Between Groups	96.466	3	32.155	1.513	.230
	Within Groups	680.243	32	21.258		
	Total	776.709	35			
磷脂酰乙醇胺	Between Groups	1728.647	3	576.216	9.412	.000
	Within Groups	1959.142	32	61.223		
	Total	3687.789	35			

表 6-45 Brown-Forsythe、Welch 均值相等检验表

		Statistic ^a	df1	df2	Sig.
溶血磷脂酰碱	Welch	6.881	3	16.072	.003
	Brown-Forsythe	16.379	3	10.892	.000
磷脂酰碱	Welch	1.638	3	17.080	.218
	Brown-Forsythe	2.539	3	20.729	.084
磷脂酰甘油	Welch	13.806	3	16.657	.000
	Brown-Forsythe	31.308	3	10.813	.000
神经鞘磷脂	Welch	1.084	3	16.540	.383
	Brown-Forsythe	1.513	3	19.979	.242
磷脂酰乙醇胺	Welch	21.419	3	14.108	.000
	Brown-Forsythe	9.412	3	19.543	.000

a. Asymptotically F distributed.

(4) 结果与讨论

在方差不齐的情形下,用表 6-44 来解释结果显然是不合适的。应使用表 6-45 来解释。

在表 6-45 中可见, 在溶血磷脂酰碱、磷脂酰甘油和磷脂酰乙醇胺指标上, 均值间无差异的原假设下, 出现目前统计量的值或更极端值的概率 (Sig.) 的值均为 0.000。而在磷脂酰碱和神经鞘磷脂指标上均值间无差异的原假设下, 出现目前统计量的值或更极端值的概率 (Sig.) 的值均大于 0.05。

(5) 结论

拒绝在溶血磷脂酰碱、磷脂酰甘油和磷脂酰乙醇胺指标上均值相等的原假设, 而认为在这三个指标的每一个上四组间至少有两个均值间存在显著性差异。本例中同 Welch 均值相等检验法的结论不矛盾。

5. 多重比较

(1) 按 Analyze→Compare Means→One-Way ANOVA 顺序, 打开 One-Way ANOVA 对话框, 见图 6-5。在变量源框中, 单击选中溶血磷脂酰碱、磷脂酰甘油和磷脂酰乙醇胺变量, 按右移箭头按钮, 将它移入到 Dependent List 下框中, 用同样的方法将实验因素变量移入到 Factor 下框中。

(2) 单击 Post Hoc 按钮, 弹出 Post Hoc Multiple Comparisons (均值差异的多重比较) 对话框 (见图 6-7)。

本例已经得到的前提条件是接受方差不齐的假设, 因此, 要进行均值差异的多重比较, 必须在图 6-7 中选择 Equal Variances Not Assumed (方差不齐) 选项。

选择 Games-Howell, 单击 Continue 按钮返回 One-Way ANOVA 对话框。

(3) 单击 OK 按钮执行, 在输出窗口中, 输出两张表, 第一张为方差分析表, 见表 6-44, 第二张为方差不齐时的多重比较, 见表 6-46。

6. 结果与讨论

关于表中各列所列表头内容, 同例 6.15 中表 6-37 的解释。根据表中两两比较时均值差值上的*号或其后的均值相等的原假设的 P 值小于 0.05 可以看出, 在溶血磷脂酰碱指标上, 间断缺氧 15d 组与对照组、急性缺氧组和间断缺氧 5d 组之间均有显著性差异, 在该指标上, 其他组之间无显著性差异。在磷脂酰甘油指标上, 间断缺氧 5d 组与对照组、急性缺氧组和间断缺氧 5d 组之间均有显著性差异, 其他组之间无显著性差异。在磷脂酰乙醇胺指标上, 同样间断缺氧 5d 组与所有组之间均有显著性差异, 其他组之间无显著性差异。

7. 结论

间断缺氧 15d 组的溶血磷脂酰碱指标值显著性地高于其他各组, 在磷脂酰甘油指标上, 间断缺氧 5d 组显著性地高于其他各组, 在磷脂酰乙醇胺指标上, 间断缺氧 5d 组显著性地低于其他各组, 其他组之间无显著性差异。

表 6-46 方差不齐时的多重比较

Games-Howell							
Dependent Variable	(I) 实验因素	(J) 实验因素	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
带血清脂酶碱	对照组	急性缺氧组	-.35000	.29953	.656	-1.2334	.5334
		间断缺氧5d组	.06556	.17381	.981	-.4413	.5724
		间断缺氧15d组	-.3.35444 [*]	.76135	.008	-5.7531	-.9558
	急性缺氧组	对照组	.35000	.29953	.656	-.5334	1.2334
		间断缺氧5d组	.41556	.27785	.475	-.4339	1.2650
		间断缺氧15d组	-.3.00444 [*]	.79160	.016	-5.4293	-.5796
	间断缺氧5d组	对照组	-.06556	.17381	.981	-.5724	.4413
		急性缺氧组	-.41556	.27785	.475	-1.2650	.4339
		间断缺氧15d组	-.3.42000 [*]	.75308	.008	-5.8146	-1.0254
	间断缺氧15d组	对照组	3.35444 [*]	.76135	.008	.9558	5.7531
		急性缺氧组	3.00444 [*]	.79160	.016	.5796	5.4293
		间断缺氧5d组	3.42000 [*]	.75308	.008	1.0254	5.8146
磷脂酰甘油	对照组	急性缺氧组	3.27667	1.66884	.255	-1.6934	8.2467
		间断缺氧5d组	-25.04222 [*]	4.86751	.002	-40.0337	-10.0508
		间断缺氧15d组	4.84444	1.75978	.068	-.2992	9.9881
	急性缺氧组	对照组	-3.27667	1.66884	.255	-8.2467	1.6934
		间断缺氧5d组	-28.31889 [*]	4.69244	.001	-43.1767	-13.4611
		间断缺氧15d组	1.56778	1.19297	.568	-1.8635	4.9990
	间断缺氧5d组	对照组	25.04222 [*]	4.86751	.002	10.0508	40.0337
		急性缺氧组	28.31889 [*]	4.69244	.001	13.4611	43.1767
		间断缺氧15d组	29.88667 [*]	4.72555	.001	15.0111	44.7622
	间断缺氧15d组	对照组	-4.84444	1.75978	.068	-9.9881	.2992
		急性缺氧组	-1.56778	1.19297	.568	-4.9990	1.8635
		间断缺氧5d组	-29.88667 [*]	4.72555	.001	-44.7622	-15.0111
磷脂酰乙醇胺	对照组	急性缺氧组	-4.89000	4.89696	.752	-18.9342	9.1542
		间断缺氧5d组	13.95000 [*]	3.22633	.010	3.7044	24.1956
		间断缺氧15d组	4.12667	3.64041	.677	-6.6333	14.8966
	急性缺氧组	对照组	4.89000	4.89696	.752	-9.1542	18.9342
		间断缺氧5d组	18.94000 [*]	3.73601	.004	6.9510	30.7290
		间断缺氧15d组	9.01667	4.09892	.181	-3.2544	21.2878
	间断缺氧5d组	对照组	-13.95000 [*]	3.22633	.010	-24.1956	-3.7044
		急性缺氧组	-18.94000 [*]	3.73601	.004	-30.7290	-6.9510
		间断缺氧15d组	-9.82333 [*]	1.79724	.002	-15.4321	-4.2146
	间断缺氧15d组	对照组	-4.12667	3.64041	.677	-14.8866	6.6333
		急性缺氧组	-9.01667	4.09892	.181	-21.2878	3.2544
		间断缺氧5d组	9.82333 [*]	1.79724	.002	4.2146	15.4321

*. The mean difference is significant at the 0.05 level.

6.3 多元正态总体均值差异的显著性检验

多元统计分析中研究的是多个随机变量之间相互依赖关系及内在的统计规律性。虽然，有时在有的问题上，将多元统计分析中考察的多个变量，一一分开考虑，也可变成一元统计分析，但由于变量多，这样做很可能会忽略变量之间的相关性，甚至会丢掉宝贵的信息。所以变量数不同，统计处理分析中所用的方法也是不尽相同的。

6.3.1 多元正态分布基本概述

1. 多元正态分布的定义

多元正态分布是一元正态分布向 $P \geq 2$ 维的推广。

设 X_1, \dots, X_p 是独立同 $N(0,1)$ 分布，则 $X=(X_1, \dots, X_p)$ 的联合概率密度为

$$f(x_1, \dots, x_p) = \left(\frac{1}{\sqrt{2\pi}} \right)^p e^{-\left(\frac{x'x}{2} \right)}$$

其中, $-\infty < x_i < \infty, i=1, 2, \dots, p, x'x = (x_1, \dots, x_p) \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} = \sum_{i=1}^p x_i^2$

称 \mathbf{X} 服从 p 元标准正态水平, 记为 $\mathbf{X} \sim N(0, I_p)$, 其中 I_p 是 p 阶单位矩阵。

2. 多元正态分布的性质

(1) 若 $(X_1, \dots, X_p)' \sim N(0, I_p)$, 则它的任意线性组合 $Y = \underset{q \times 1}{A} \underset{q \times p}{X} + \underset{q \times 1}{\mu}$ 仍服从多元正态分布, 且 $EY = \mu, \text{Var}(Y) = E(Y - EY)(Y - EY)' = AA'$, 从而 $Y \sim N(\mu, AA')$ 。

(2) 设 X 是 p 维随机向量, 则 $X \sim N_p(\mu, \Sigma)$ 的充分必要条件为其任一线性函数 $a'X = a_1X_1 + a_2X_2 + \dots + a_pX_p, a \in R^p$ 服从 $N(a'\mu, a'\Sigma a)$ 分布。

(3) 若 $X \sim N_p(\mu, \Sigma)$, 常数矩阵 $\underset{q \times p}{A} = \begin{bmatrix} a_{11}, \dots, a_{1p} \\ \vdots \\ a_{q1}, \dots, a_{qp} \end{bmatrix}$, 则 $Y = AX$ 服从 $N_q(A\mu, A\Sigma A')$ 分布。

进一步, 有 $Y = AX + b$ 服从 $N_q(A\mu + b, A\Sigma A')$, 其中, $b \in R^q$,

(4) 若 $Y = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_{(1)} \\ \mu_{(2)} \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right), i=1, 2$, 则 $Y_{(1)}, Y_{(2)}$ 相互独立的充分必要条件是: $\Sigma_{12} = 0$ 。

3. 多元正态总体样本的数字特征

(1) 样本均值向量

$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = (\bar{X}_1, \dots, \bar{X}_p)'_{p \times 1}$, 其中 $\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}, i=1, 2, \dots, p$

(2) 样本离差矩阵

称 $S = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})' = (S_{ij})_{p \times p}$ 为样本离差矩阵。

其中 $S_{ij} = \sum_{a=1}^n (X_{ai} - \bar{X}_i)(X_{aj} - \bar{X}_j), (i, j=1, 2, \dots, p)$ 。

(3) 样本协方差矩阵

称 $R = \frac{1}{n-1} S$ 作为样本协方差矩阵。

(4) 样本相关矩阵

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ & & \cdots & \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

其中 $\rho_{ij} = \rho_{ji} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}, (i, j = 1, 2, \cdots, p)$

(5) 使用多元正态总体均值差异的显著性检验方法时的条件

在进行多元多总体的均值差异的比较时, 必须满足以下的一些的条件:

① 观察独立。

这由试验设计决定, 也能有此知道观察是否独立。

② 多元总体必须服从多元正态。

用 SPSS 直接进行多元总体的正态性检验是困难的。多元正态有三个方面的假定: 一是各个变量正态, 二是变量的线性组合应是正态的, 三是所有变量的子集有多元正态。根据上文提到的多元正态总体的性质 (1) 可知, 只要各个样本的每个指标的数据值都服从标准正态分布, 则其多元总体服从多元正态分布。

③ 总体协方差阵相等或方差齐性。

④ 随机误差独立。

这一点在做多元多正态总体均数比较时, 不是强制性的, 即多元变量间可以有低到中度的相关, 可以配合相关分析, 查看相关系数值是否小于 0.5 来判定。

6.3.2 多元正态总体均值差异的检验方法

设 L 是一个 $l \times p$ 的已知矩阵, M 是一个 $r \times m$ 的已知矩阵以及 K 是一个 $l \times m$ 的已知矩阵。当且仅当 LB 可以估计时, 检验假设 $H_0: LBM = K$ 与备择假设 $H_1: LBM \neq K$ 是可以检验的。

假设 SSCP 矩阵为: $S_H = (L\hat{B}M - K)'(LGL')^{-1}(L\hat{B}M - K)$, 误差 SSCP 矩阵为: $S_E = M'SM$ 。可以使用基于 $S_E^{-1}S_H$ 的特征值的四种检验统计量: Wilks' lambda, Hotelling-Lawley trace, Pillai's trace, 以及 Roy's 最大特征值。

设 $S_E^{-1}S_H$ 的特征值是 $\lambda_1 \geq \cdots \geq \lambda_{r_E} \geq 0$ 且 $\lambda_{r_E+1}, \cdots, \lambda_m = 0$, 并设 $r_E = \text{rank}(S_E)$, $s = \min(l, r_E)$, $n_e = n - r_x$, $m^* = \frac{1}{2}(|r_E - l| - 1)$, $n^* = \frac{1}{2}(n_e - r_E - 1)$ 。

1. Wilks' lambda (λ) 检验法

$\Lambda = \frac{|S_E|}{|S_H + S_E|} = \prod_{k=1}^m \frac{1}{(1 + \lambda_k)}$, 当 H_0 为真时, F 统计量: $F = \frac{(\zeta\tau - 2\nu)(1 - \Lambda^{1/\tau})}{lr_E(\Lambda^{1/\tau})}$ 近似

服从 F 分布, 这里: $\varsigma = n_e - \frac{1}{2}(r_E - l + 1)$, $\nu = \frac{1}{4}(lr_E - 2)$

$$\tau = \begin{cases} \sqrt{(l^2 r_E^2 - 4)/(l^2 + r_E^2 - 5)} & \text{如果 } l^2 + r_E^2 - 5 > 0 \\ 1 & \text{其他} \end{cases}$$

自由度为 $(lr_E, \varsigma\tau - 2\nu)$ 。如果 $s=1, 2$, 则 F 统计量是精确的。

η^2 统计量为: $\eta^2 = 1 - \Lambda^{1/s}$ 。非中心性参数为 $\lambda = (\xi\tau - 2\nu)\eta^2 / (1 - \eta^2)$ 。

功效为 $1 - NCDF.F(F_\alpha, lr_E, (\xi\tau - 2\nu), \lambda)$, 这里的 F_α 是中心 F 分布的上 100α 百分点, 并且 α 是用户在 CRITERIA 子命令中的 ALPHA 关键词上指定。

2. Hotelling-Lawley 迹检验法

在 SPSS 中, Hotelling-Lawley trace 被缩写成 Hotelling's trace。

$$T = \text{trace}(\mathbf{S}_E^{-1} - \mathbf{S}_H) = \sum_{k=1}^m \lambda_k。$$

当 H_0 为真时, F 统计量为: $F = \frac{2(sn^* + 1)T}{s(2m^* + s + 1)s}$, 它近似服从自由度为

$(s(2m^* + s + 1), 2(sn^* + 1))$ 的 F 分布。如果 $S=1$, 则 F 统计量是精确的。

η^2 统计量为: $\eta^2 = (T/s)/(T/s + 1)$ 。非中心性参数为 $\lambda = 2(sn^* + 1)\eta^2 / (1 - \eta^2)$ 。

功效为 $1 - NCDF.F(F_\alpha, s(2m^* + s + 1), 2(sn^* + 1), \lambda)$, 这里的 F_α 是中心 F 分布的上 100α 百分点, 并且 α 是用户在 CRITERIA 子命令中的 ALPHA 关键词上指定。

3. Pillai 迹检验法

$$V = \text{trace}(\mathbf{S}_H(\mathbf{S}_H + \mathbf{S}_E)^{-1}) = \sum_{k=1}^m \lambda_k / (1 + \lambda_k)。$$

当 H_0 为真时, F 统计量为: $F = \frac{(2n^* + s + 1)V}{(2m^* + s + 1)(s - V)}$, 它近似服从自由度为

$(s(2m^* + s + 1), s(2n^* + s + 1))$ 的 F 分布。如果 $S=1$, 则 F 统计量是精确的。

η^2 统计量为: $\eta^2 = V/s$ 。非中心性参数为 $\lambda = s(2n^* + s + 1)\eta^2 / (1 - \eta^2)$ 。

功效为 $1 - NCDF.F(F_\alpha, s(2m^* + s + 1), s(2n^* + s + 1), \lambda)$, 这里的 F_α 是中心 F 分布的上 100α 百分点, 并且 α 是用户在 CRITERIA 子命令中的 ALPHA 关键词上指定。

4. Roy 最大特征值

$\Theta = \lambda_1$, 它是 $\mathbf{S}_E^{-1}\mathbf{S}_H$ 的最大的特征值。

当 H_0 为真时, F 统计量为: $F = \Theta(n_e - \omega + r_H)/\omega$, 这里, $\omega = \max(1, r_E)$ 是 F 的上限, 它产生一个显著性水平的下限。自由度是 $\omega, n_e - \omega + r_H$ 。如果 $S=1$, 则 F 统计量是精

确的。

η^2 统计量为: $\eta^2 = \Theta / (1 + \Theta)$ 。非中心性参数为 $\lambda = (n_e - \omega + r_H) \eta^2 / (1 - \eta^2)$ 。

功效为 $1 - \text{NCDF}.F(F_\alpha, \omega, n_e - \omega + 1, \lambda)$, 这里的 F_α 是中心 F 分布的上 100α 百分点, 并且 α 是用户在 CRITERIA 子命令中的 ALPHA 关键词上指定。

6.3.3 多个协方差阵相等检验—Box's M 检验

Box (1949 年) 根据拟然比检验得到一个检验统计量。该检验统计量被称为 Box's M 统计量。Box's M 统计量通常用于协方差矩阵的齐性检验。适合于样本容量中到小的样本, F 的近似值用来计算它的显著性。

记 r 个因变量中第 j 个因变量在第 i 个水平的观测是: $y'_{ij} = x'_{ij} \mathbf{B} + e'_{ij}$ 。其中, $e_{ij} \sim w_{ij}^{-1} \sum_i$, $i=1, \dots, g, j=1, \dots, n_i$

检验协方差矩阵齐性的原假设: $H_0: \Sigma_1 = \dots = \Sigma_g$

$$\text{Box's M 统计量: } M = \begin{cases} (n-g) \log |\mathbf{S}| - \sum_{i=1}^g (n_i-1) \log |\mathbf{S}_i| & \text{if } |\mathbf{S}| > 0 \\ 0 & \text{if } |\mathbf{S}| \leq 0 \end{cases}$$

其中,

$$\text{单元协方差矩阵: } S_i = \begin{cases} \sum_{j=1}^{n_i} w_{ij} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i) (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)' / (n_i - 1) & \text{if } n_i > 1 \\ 0 & \text{if } n_i \leq 1 \end{cases}$$

$$\text{合并协方差矩阵: } S = \begin{cases} \sum_{i=1}^g (n_i - 1) \mathbf{S}_i / (n - g) & \text{if } n > g \\ 0 & \text{if } n \leq g \end{cases}$$

平均数: $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$ 。

g : 非奇异协方差矩阵的单元数。

n_i : 在第 i 个单元的样品数。

n : 总的样本含量, $n = n_1 + \dots + n_g$

y_{ij} : 第 j 个因变量在第 i 单元。长度 r 的列向量。

w_{ij} : 同 y_{ij} 有关的回归权重。假定 $w_{ij} > 0$ 。

显著性: $1 - \text{CDF}.F(\gamma M, f_1, f_2)$, CDF.F 是用来计算 F 分布的 SPSS 函数。其中:

$$f_1 = (g-1)r(r+1)/2$$

$$\rho = 1 - \frac{2r^2 + 3r - 1}{6(r+1)(g+1)} \left(\sum_{i=1}^g \frac{1}{(n_i-1)} - \frac{1}{(n-g)} \right)$$

$$\tau = \frac{(r-1)(r+2)}{6(g-1)} \left(\sum_{i=1}^g \frac{1}{(n_i-1)^2} - \frac{1}{(n-g)^2} \right)$$

$$f_2 = \frac{f_1 + 2}{\left| \tau - (1-\rho)^2 \right|}$$

$$\gamma = \frac{(\rho - f_1/f_2)}{f_1}$$

6.3.4 随机误差的独立性检验—Bartlett 球型检验

在做多元正态总体检验时，只有在任选项中要求输出残差 SSCP 矩阵时，才会在输出窗口中，输出 Bartlett 球型检验结果。

在 Bartlett 球型检验中原假设是 $H_0: \Sigma = \sigma^2 \mathbf{I}_r$ ，对应的备择假设为： $H_1: \Sigma \neq \sigma^2 \mathbf{I}_r$ ，这里， $\sigma^2 > 0$ ，没作说明，并且 \mathbf{I}_r 是 $r \times r$ 的单位矩阵。

似然比检验统计量为

$$\lambda = \begin{cases} \frac{|\mathbf{A}|^{n/2}}{(\text{trace}(\mathbf{A})/r)^{nr/2}} & \text{if } \text{trace}(\mathbf{A}) > 0 \\ \text{SYSMIS} & \text{if } \text{trace}(\mathbf{A}) \leq 0 \end{cases}$$

这里 $\mathbf{A} = (\mathbf{Y} - \mathbf{XB})' \mathbf{W} (\mathbf{Y} - \mathbf{XB})$ 是 $r \times r$ 的残差平方和矩阵和向量积。

χ^2 近似值：取 $W = \lambda^{2/n}$ 。当 n 较大，且在原假设下，对于 $n - r_{\bar{x}} \geq 1$ 且 $r \geq 2$ ，

$$p_r(-\rho(n - r_{\bar{x}}) \log W \leq c) = p_r(\chi_f^2 \leq c) + \omega_2(p_r(\chi_{f+4}^2 \leq c) - p_r(\chi_f^2 \leq c)) + O(n^{-3})$$

这里， $f = r(r+1)/2 - 1$ ， $\rho = 1 - (2r^2 + r + 2)/(6r(n - r_{\bar{x}}))$

$$\omega_2 = \frac{(r+2)(r-1)(r-2)(2r^3 + 6r^2 + 3r + 2)}{288r^2(n - r_{\bar{x}})^2 \rho^2}$$

$$\chi^2 \text{ 统计量: } C = \begin{cases} -\rho(n - r_{\bar{x}}) \log W & \text{if } W > 0 \\ \text{SYSMIS} & \text{其他} \end{cases}$$

自由度： $f = r(r+1)/2 - 1$ 。

显著性： $1 - \text{CDF.CHISQ}(C, f) - \omega_2(\text{CDF.CHISQ}(C, f+4) - \text{CDF.CHISQ}(C, f))$ 。

这里的 CDF.CHISQ 是累积 χ^2 分布的 SPSS 函数，由于浮点精度的原因，只要计算值小于 0，显著性被设置为 0。

6.3.5 实例分析

1. 多元正态总体均值等于常数向量的检验

在许多实际的科学研究中，通常会遇到对不同总体的多个指标需要同时进行考察的情况，此时，可以用 p 项指标的历史记录的平均值作为 μ_0 ，探寻新的 p 项指标的平均值是否与历史记录的平均值间有差异。

例 6.19 在研究排汗量与体内钠含量与钾含量关系的研究中，测试了 20 名健康女性在这三个指标上的具体数据，见例 2.10 表 2-7，数据存放在 data06-11.sav 中。试分析这三个指标的均值是否是 4、50、10？

在 SPSS 中，检验此类问题的操作步骤如下：

1. 在数据编辑窗口，打开 data06-11.sav。
2. 对排汗量、钠含量、钾含量的数据资料进行正态性检验
用第 2 章 2.4 节探索分析中介绍的步骤，可得表 6-47。

由表 6-47 中 P 值都大于 0.05 可知，现有证据不足以推翻这三个变量服从正态分布的原假设。

表 6-47 数据资料的正态性检验

Tests of Normality						
类别	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
排汗量	.159	20	.197	.976	20	.869
钠含量	.101	20	.200 [*]	.986	20	.986
钾含量	.129	20	.200 [*]	.964	20	.623

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

3. 计算排汗量、钠含量、钾含量分别与其已知总体均值 4、50、10 的差值。

利用 Transform 菜单中的 Compute Variable 过程，按第 3 章 3.3 节计算派生指标中介绍的方法，计算得到排汗量转换、钠含量转换、钾含量转换值，见图 6-9。并将转换结果另存为 data06-11a.sav。

类别	排汗量	钠含量	钾含量	排汗量转换	钠含量转换	钾含量转换
1	3.70	48.50	9.30	-0.30	-1.50	-0.70
2	5.70	66.10	8.00	1.70	16.10	-2.00
3	3.80	47.20	10.90	-0.20	-2.80	0.90
4	3.20	53.20	12.00	-0.80	3.20	2.00
5	3.10	66.50	9.70	-0.90	5.50	-0.30
6	4.60	36.10	7.90	0.60	-13.90	-2.10
7	2.40	24.80	14.00	-1.60	-26.20	4.00
8	7.20	33.10	7.60	3.20	-16.90	-2.40
9	6.70	47.40	8.50	2.70	-2.60	-1.50
10	5.40	54.10	11.30	1.40	4.10	1.30
11	3.90	36.90	12.70	-0.10	-13.10	2.70
12	4.50	58.00	12.30	0.50	8.00	2.30
13	3.50	27.00	9.80	-0.50	-22.20	-0.20
14	4.50	40.20	8.40	0.50	-9.80	-1.60
15	1.50	13.50	10.10	-2.50	-36.50	0.10
16	8.50	56.40	7.10	4.50	6.40	-2.90
17	4.50	71.60	8.20	0.50	21.60	-1.80
18	6.50	52.00	10.90	2.50	2.00	0.90
19	4.10	44.10	11.20	0.10	-5.90	1.20
20	5.50	40.90	9.40	1.50	-9.10	-0.60

图 6-9 计算转换值

注：排汗量转换=排汗量-4，钠含量转换=钠含量-50，钾含量转换=钾含量-10，这样做的目的是，当排汗量、钠含量、钾含量分别与其已知总体均值 4、50、10 间无差异时，就等价于要检验它们的差值都与已知均值为 0 的总体间无差异。

4. 检验

(1) 按 Analyze → General Linear Model → GLM Multivariate 顺序，打开 GLM Multivariate 对话框，见图 6-10。

在左侧的变量源框中，选中排汗量转换、钠含量转换、钾含量转换变量，单击右移箭头按钮，将这三个变量移入到 Dependent Variables 框中（图 6-10）。

(2) 单击 Model 按钮，打开 Model 对话框，见图 6-11。关闭 Include intercept in model，其他不作任何选择。单击 Continue 按钮，返回 GLM Multivariate 对话框。

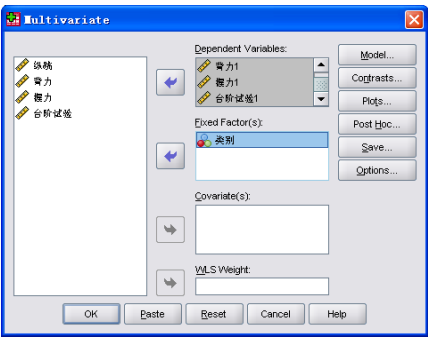


图 6-10 GLM Multivariate 对话框

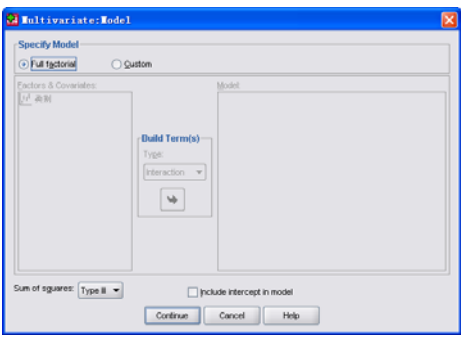


图 6-11 Model 对话框

(3) 单击 OK 按钮执行，在输出窗口中，得到二张表，第一张表检验结果表，见表 6-48。第二张组间效应检验表，在本类题型中无效，不作介绍。

表 6-48 多元正态总体均值等于常数向量的检验

Multivariate Tests ^a						
Effect		Value	F	Hypothesis df	Error df	Sig.
类别	Pillai's Trace	.339	2.905 ^a	3.000	17.000	.065
	Wilks' Lambda	.661	2.905 ^a	3.000	17.000	.065
	Hotelling's Trace	.513	2.905 ^a	3.000	17.000	.065
	Roy's Largest Root	.513	2.905 ^a	3.000	17.000	.065

a. Exact statistic
b. Design: 类别

5. 结果与讨论

表 6-48 显示，4 种检验方法的统计量 F 值为 2.905，第一自由度为 3，第二自由度为 17， P 均为 0.065。

6. 结论

因为 $P > 0.05$ ，所以没有足够的证据拒绝这三个变量的转换值均值等于 0 向量的原假设。在 0.05 水平上，目前只有认为这三个指标的均值是 4、50、10。

2. 多元两正态总体均值的检验

例 6.20 为了研究美、日两国在华投资企业对中国经营环境的评价是否存在差异，从日、美两国在华投资企业中随机各抽取 10 家企业，让他们对中国的政治、经济、法律、文化等环境评分，得表 6-49，试问两国企业对中国经营环境的评价有无差异？资料来源：国务院发展研究中心 APEC 在华投资企业情况调查。

要用上述方法进行检验的一个前提条件是：各个指标的数据资料必须服从正态分布，且各组之间的协方差矩阵相等。

在 SPSS16.0 中，检验此类问题的操作步骤如下：

1. 在 SPSS 数据编辑窗口中，将表 6-49 中的数据建立成数据文件，见 data06-12.sav。
 2. 对政治环境、经济环境、法律环境和文化环境的数据资料进行正态性检验。
- 仿上例做法，可得正态性检验结果，见表 6-50。

表 6-49 美、日企业对中国经营环境的评价

企业类别		政治环境	经济环境	法律环境	文化环境
美国企业	1	65.00	35.00	25.00	60.00
	2	75.00	50.00	20.00	55.00
	3	60.00	45.00	35.00	65.00
	4	75.00	40.00	40.00	70.00
	5	70.00	30.00	30.00	50.00
	6	55.00	40.00	35.00	65.00
	7	60.00	45.00	30.00	60.00
	8	65.00	40.00	25.00	60.00
	9	60.00	50.00	30.00	70.00
	10	35.00	55.00	35.00	75.00
日本企业	1	55.00	55.00	40.00	65.00
	2	50.00	60.00	45.00	70.00
	3	45.00	45.00	35.00	75.00
	4	50.00	50.00	50.00	70.00
	5	55.00	50.00	30.00	75.00
	6	60.00	40.00	45.00	60.00
	7	65.00	55.00	45.00	75.00
	8	50.00	65.00	35.00	80.00
	9	40.00	45.00	30.00	65.00
	10	45.00	50.00	45.00	70.00

表 6-50 正态检验结果

Tests of Normality						
企业类别	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
政治环境 美国企业	.206	10	.200 [*]	.901	10	.225
政治环境 日本企业	.180	10	.200 [*]	.966	10	.854
经济环境 美国企业	.155	10	.200 [*]	.969	10	.886
经济环境 日本企业	.160	10	.200 [*]	.942	10	.575
法律环境 美国企业	.174	10	.200 [*]	.952	10	.691
法律环境 日本企业	.260	10	.053	.887	10	.158
文化环境 美国企业	.155	10	.200 [*]	.969	10	.886
文化环境 日本企业	.174	10	.200 [*]	.952	10	.691

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

从表 6-50 可见，在各变量服从正态分布的原假设下，出现目前统计量的值或更极端值的概率（Sig.）的值均大于 0.05。所以，现有证据不足以推翻原假设。故不能拒绝数据资料服从正态分布的假设。

3. 检验

按 Analyze→General Linear Model→GLM Multivariate 顺序，打开 GLM Multivariate 对话框，见图 6-10。

在左侧的变量源框中，选中对政治环境、经济环境、法律环境和文化环境变量，按右移箭头按钮，将这三个变量移入到 Dependent Variables 框中（见图 6-10）。选中组别变量将其移到 Fixed Factor(s)框中。

单击 Model 按钮，打开 Model 对话框，见图 6-11。关闭 Include intercept in model，其他不作任何选择。单击 Continue 按钮，返回 GLM Multivariate 对话框。

单击 Option 按钮，展开 Option 对话框，见图 6-12。在 Display 中选择 Homogeneity tests 选项，要求做方差的齐性检验。选择 Residual SSCP matrix 选项，要求输出 Bartlett 球型检验结果。

单击 Continue 按钮，返回 GLM Multivariate 对话框。

单击 OK 按钮执行，在输出窗口中，出现输出结果，见表 6-51、表 6-52、表 6-53、表 6-54、表 6-55、表 6-56、表 6-57。



图 6-12 Option 对话框

表 6-51 组间因素

Between-Subjects Factors		
	Value Label	N
企业类别	1 美国企业	10
	2 日本企业	10

表 6-52 协方差矩阵的齐性检验

Box's Test of Equality of Covariance Matrices ^a	
Box's M	12.528
F	.945
df1	10
df2	1.549E3
Sig.	.490
Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.	
a. Design: 企业类别	

表 6-53 Bartlett 球型检验结果

Bartlett's Test of Sphericity ^a	
Likelihood Ratio	.003
Approx. Chi-Square	9.485
df	9
Sig.	.396
Tests the null hypothesis that the residual covariance matrix is proportional to an identity matrix.	
a. Design: 企业类别	

表 6-54 方差的齐性检验

Levene's Test of Equality of Error Variances ^a				
	F	df1	df2	Sig.
政治环境	.012	1	18	.913
经济环境	.214	1	18	.649
法律环境	.878	1	18	.361
文化环境	.676	1	18	.422
Tests the null hypothesis that the error variance of the dependent variable is equal across groups.				
a. Design: 企业类别				

表 6-55 多元方差分析检验

Multivariate Tests ^a						
Effect		Value	F	Hypothesis df	Error df	Sig.
企业类别	Pillai's Trace	1.620	17.029	8.000	32.000	.000
	Wilks' Lambda	.002	92.199 ^b	8.000	30.000	.000
	Hotelling's Trace	247.049	432.336	8.000	28.000	.000
	Roy's Largest Root	245.392	9.816E2 ^b	4.000	16.000	.000

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: 企业类别

表 6-56 组间效应检验

Tests of Between-Subjects Effects						
Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	政治环境	67482.500 ^a	2	33741.250	611.932	.000
	经济环境	44500.000 ^a	2	22250.000	445.000	.000
	法律环境	25302.500 ^a	2	12651.250	294.786	.000
	文化环境	89392.500 ^d	2	44696.250	966.405	.000
企业类别	政治环境	67482.500	2	33741.250	611.932	.000
	经济环境	44500.000	2	22250.000	445.000	.000
	法律环境	25302.500	2	12651.250	294.786	.000
	文化环境	89392.500	2	44696.250	966.405	.000
Error	政治环境	982.500	18	55.139		
	经济环境	900.000	18	50.000		
	法律环境	772.500	18	42.917		
	文化环境	832.500	18	46.250		
Total	政治环境	68475.000	20			
	经济环境	45400.000	20			
	法律环境	26075.000	20			
	文化环境	90225.000	20			

a. R Squared = .986 (Adjusted R Squared = .984)

b. R Squared = .980 (Adjusted R Squared = .978)

c. R Squared = .970 (Adjusted R Squared = .967)

d. R Squared = .991 (Adjusted R Squared = .990)

表 6-57 残差 SSCP 矩阵

Residual SSCP Matrix				
	政治环境	经济环境	法律环境	文化环境
Sum-of-Squares and Cross-Products	政治环境	992.500	-110.000	55.000
	经济环境	-110.000	900.000	60.000
	法律环境	55.000	60.000	772.500
	文化环境	-252.500	505.000	160.000
Covariance	政治环境	55.139	-6.111	3.056
	经济环境	-6.111	50.000	3.333
	法律环境	3.056	3.333	42.917
	文化环境	-14.028	28.056	8.889
Correlation	政治环境	1.000	-.116	.063
	经济环境	-.116	1.000	.072
	法律环境	.063	.072	1.000
	文化环境	-.278	.583	.200

Based on Type III Sum of Squares

4. 结果与讨论

表 6-51 列出了组间因素为美国企业和日本企业。

表 6-52 列出的是 Box 检验结果，它表明在各组之间协方差矩阵相等的原假设下，出现目前统计量的值或更极端值的概率（Sig.）的值为 0.490 大于 0.05，所以，不拒绝原假设。

表 6-53 列出了 Bartlett 球型检验结果，由于在随机误差是独立的原假设下，出现目前统计量的值或更极端值的概率（Sig.）的值为 0.396 大于 0.05，故不拒绝原假设，即认为随机误差是独立的。

表 6-54 列出了等方差检验结果，由于在方差齐性的原假设下，出现目前统计量的值或更极端值的概率（Sig.）的值大于 0.05。说明方差是齐性的。

表 6-55 列出了多元方差检验结果，从 Hotelling's trace 等四种检验的结果可见，在总体均数间无差异的原假设下，出现目前统计量的值或更极端值的概率（Sig.）的值为 0.000，小于 0.05，故拒绝原假设。

表 6-56 列出了组间效应检验，由于在组间均数间无差异的原假设下，出现目前统计

量的值或更极端值的概率 (Sig.) 的值都为 0.000 均小于 0.01, 故两组在四个指标上均有极显著性差异。

表 6-57 列出了残差 SSCP 矩阵, 它包括平方和和交叉乘积矩阵、协方差矩阵和相关系数矩阵。

5. 结论

由于各指标的数据服从正态, 又不能拒绝在各组之间的协方差矩阵是相等的原假设, 且方差齐性, 随机误差是独立的, 因而, 其线性组合服从多元正态, 所以可用多元方差分析进行多元两正态总体均值间差异的显著性检验, 且检验结果显著, 说明两总体之间是有差异的。组间效应检验表明, 两组在各个指标上均有极显著性差异。

6.4 非正态总体参数的假设检验

在以上求正态总体均值差异的显著性检验中, 找到的统计量都有精确的分布, 因此, 对样本容量没有任何限制, 但对于大多数非正态总体, 一般很难找到具有精确分布的统计量来进行检验, 故通常需要采用大样本 ($n \geq 100$) 的方法, 根据中心极限定理, 利用极限分布对总体参数作近似检验, 也即用近似分布为正态分布的统计量作为检验统计量。

6.4.1 非正态总体的均值检验

6.4.1.1 关于对一个已知总体均值 μ_0 的假设检验

从非正态总体 X 中随机抽取一个大样本 (X_1, X_2, \dots, X_n) , 当非正态总体的方差未知, 需要考虑 $H_0: \mu = \mu_0$ 时, 可用 $U = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ 作为统计量, 由于样本容量一般达 100 以上, 因此, 可以证明 $U \sim N(0,1)$ 。

在有样本原始数据的前提下, 由于对应的 $t_{\alpha(n)}$ 值小于 u_α 值, 因此, 对于不提高犯第一类错误的概率的前提下, 也可直接用 6.2.1 中的单样本 t 检验来做近似检验。

而在已知非正态总体样本统计量的前提下, 可用下述方法进行检验。

例 6.21 对木材等级的划分可用其小头直径与 12cm 的标准进行比较来决定, 当其平均直径大于 12cm 时, 称该批木材为一等品。现从一批木材中随机抽取 120 根, 测得其平均直径为 $\bar{X} = 12.7\text{cm}$, 标准差为 $S = 2.8\text{cm}$, 问能否认为这批木材属于一等品?

本例所作的原假设为: $H_0: \mu \leq \mu_0 = 12$, 备择假设为: $H_1: \mu > \mu_0$ 。

在 SPSS 中的解题步骤为:

(1) 在数据编辑窗口中建立如图 6-13 所示的数据文件 data06-13.sav。

(2) 按 Transform → Compute Variable 顺序, 打开 Compute Variable 主对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为原假设成立的概率值。在 Numeric

Expression 框中, 输入 “(1-CDF.NORMAL((样本均值-已知均值)/样本标准差*sqrt(样本容量), 0, 1))”。

如果本例的原假设改为双侧检验即 $H_0: \mu = \mu_0$, 则在 Numeric Expression 框中, 输入 “(1-CDF.NORMAL(abs((样本均值-已知均值)/样本标准差*sqrt(样本容量)),0,1))*2”。

(3) 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现原假设成立的概率值的新变量及其值。见图 6-14。

样本均值	样本标准差	样本容量	已知均值
12.70	2.80	120.00	12.00

图 6-13 数据文件中的内容

样本均值	样本标准差	样本容量	已知均值	原假设成立的概率值
12.70	2.80	120.00	12.00	0.00

图 6-14 U 检验结果

6.4.1.2 关于对一个已知总体率 p_0 的假设检验

设 $X = \begin{cases} 1, A \\ 0, \bar{A} \end{cases}$, 事件 A 发生率为 p , 则总体 $X \sim b(1, p)$, 现从该总体中抽取一个大样

本, 要检验假设 $H_0: p = p_0$, 由于 p 的 MLE 为 $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$, $E(\bar{X}) = p$, $D(\bar{X}) = \frac{p(1-p)}{n}$,

根据中心极限定理可知, $\frac{\bar{X} - p}{\sqrt{p(1-p)/n}} = \frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}}$ 的极限分布为 $N(0,1)$ 。故当 H_0 成立时,

检验统计量 $U = \frac{\bar{X} - p_0}{\sqrt{p(1-p)/n}} = \frac{p - p_0}{\sqrt{p(1-p)/n}} \sim N(0,1)$ 。当 $P(H_0) < \alpha$, 拒绝 H_0 。

当样本含量较小时, 则采用二项分布精确检验法, 它为非参数假设检验。见 7.1。

例 6.22 某厂产品的合格率为 95%, 现从其生产的一批产品中, 随机抽取 100 件, 结果有 6 件不合格, 能否认为该厂生产的这批产品没有达到合格率的要求?

在 SPSS 中的解题步骤如下:

(1) 在数据编辑窗口中, 建立数据文件, 见 data06-14.sav。

(2) 加权处理

按 Data→Weight Cases 顺序, 打开 Weight Cases 对话框, 见图 2-73。选择 Weight Cases by 选项, 在左侧变量名源框中, 选中 观察数变量, 并按右移箭头将 观察数移入 Frequency Variable 框中, 作为加权变量。单击 OK 按钮, 完成加权设置。

(3) 按 Analyze→Nonparametric Tests→Binomial 顺序, 打开 Binomial Test 对话框, 见图 6-15。

在左侧变量名列表中选择合格情况变量将其移入到 Test Variable List 下框中。

在 Test Proportion 后框中输入 0.95。其他保持系统默认选项，即采用渐近分布检验法。

(4) 单击 OK 按钮运行，则在输出窗口中，得到计算结果，见表 6-58。

(5) 结果与讨论

表 6-58 从左到右依次列出的是变量名(合格情况)、事件类别(合格、不合格)、实际观察数(94、6)、观察概率(0.94、0.06)、检验概率(0.95)、出现目前统计量的值或更极端值的概率(Sig.)的值的渐近分布单侧概率(0.384)。

备注 a 指出备择假设陈述第一组样品的比例 <0.95 。这说明在本程序中，采用的是单侧检验，原假设为第一组样品的比例大于等于 0.95。

备注 b 指出，现有的结果是根据 Z 的渐近分布计算的。

(6) 结论

由于在第一组样品的比例大于等于 0.95 的概率的原假设下，出现目前统计量的值或更极端值的概率(Sig.)的值为 0.384 大于 0.05，故不拒绝原假设，即认为不能认为该厂生产的这批产品没有达到合格率的要求。

注意，本例采用的渐近分布的 Z 检验，故是参数性假设检验。

6.4.1.3 两个二项分布总体参数的检验

设 X_1, X_2, \dots, X_{n_1} 与 Y_1, Y_2, \dots, Y_{n_2} 是分别从总体 $X \sim B(1, p_1), Y \sim B(1, p_2)$ 中随机抽取的两个相互独立的样本，且 $n_1 p_1 > 5, n_1(1-p_1) > 5, n_2 p_2 > 5, n_2(1-p_2) > 5$ ，因为

$$\bar{X} \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \bar{Y} \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

所以

$$\bar{X} - \bar{Y} \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

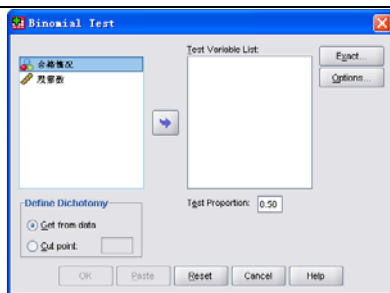


图 6-15 Binomial Test 对话框

表 6-58 一个已知总体率 p_0 的假设检验结果

Binomial Test					
	Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
合格情况 Group 1	合格	94	.94	.95	.384 ^{a,b}
Group 2	不合格	6	.06		
Total		100	1.00		

a. Alternative hypothesis states that the proportion of cases in the first group $< .95$.

b. Based on Z Approximation.

故在 $H_0: \bar{X} = \bar{Y}$ 成立的前提下, $U = \frac{\bar{X} - \bar{Y} - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$, 当 $P(H_0) < \alpha$ 时,

拒绝原假设。

例 6.23 为确定肥料的效果, 取 1000 株植物做试验, 其中在 900 株施肥的植物中有 783 株长势良好, 而在 100 株未施肥的植物中有 53 株长势良好, 问施肥是否有效 ($\alpha = 0.05$) ?

本例的原假设为: $H_0: \bar{X} = \bar{Y}$, 也即两种方法下长势的良好率相等。

在 SPSS 中的解题步骤如下:

(1) 在数据编辑窗口中建立如图 6-16 所示的数据文件 data06-15.sav。

(2) 按 Transform→Compute Variable 顺序, 打开 Compute Variable 主对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为原假设成立的概率值。在 Numeric Expression 框中, 输入 “(1-CDF.NORMAL(abs(良好 1/施肥-良好 2/未施肥)/sqrt(良好 1/施肥*(1-良好 1/施肥)/施肥+良好 2/未施肥*(1-良好 2/未施肥)/未施肥),0,1))*2”。

注: 上式等价于输入

“(1-CDF.NORMAL(abs(783/900-53/100)/sqrt(783/900*(1-783/900)/900+53/100*(1-53/100)/100),0,1))*2”。

(3) 单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现原假设成立的概率值的新变量及其值。见图 6-17。

施肥	良好1	未施肥	良好2
900.00	783.00	100.00	53.00

图 6-16 数据文件中的内容

施肥	良好1	未施肥	良好2	原假设成立的概率
900.00	783.00	100.00	53.00	0.0000

图 6-17 两率差异性显著性检验结果

(4) 结论

由于在两种方法下长势的良好率相等的原假设下, 出现目前统计量的值或更极端值的概率 (Sig.) 的值为 0.0000, 因此, 拒绝原假设, 认为两种方法下长势的良好率是不一样的, 由于施肥植物中的长势良好率高于未施肥植物中的长势良好率, 因此, 可以认为施肥是有效的。

6.4.2 指数分布总体参数的检验

设总体 X 服从指数分布, 其概率密度函数为 $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$, 从指数分布的总体

X 中随机抽取一个样本 (X_1, X_2, \dots, X_n) 。

检验假设 $H_0: \lambda = \lambda_0$

检验统计量: $\chi^2 = 2n\lambda\bar{X} = 2\lambda\sum_{i=1}^n X_i \sim \chi^2(2n)$

式中 χ^2 的中文名为卡方, 它不是 x^2 。

当 $P(H_0) < 0.05$, 拒绝原假设。

例 6.24 已知每次前后相邻的两辆车到达路口的时间间隔服从指数分布且相互独立, 其中参数 λ 为平均每秒的车流量。现观察到各车辆到达路口的时间间隔 (单位: 秒), 并将得到的数据资料存放在数据文件 data06-15.sav 中。可否认为 $\lambda = 0.6$?

(1) 在数据编辑窗口中, 打开数据文件 data06-15.sav。

(2) 按 Descriptive Statistics→Descriptive 顺序, 展开 Descriptive 对话框。

将左侧变量名源中的 *时间间隔* 变量移入到 Variable[s] 下框中。

(3) 单击 Options 按钮, 在弹出的 Options 选项卡中, 只选择 Sum 选项, 其他处于关闭状态。单击 Continue 按钮返回 Descriptive 对话框。

(4) 单击 OK 按钮运行, 则在输出窗口中, 得到表 6-59, 得到 34 个观察数据的总和为 55.90。

(5) 作假设检验。

$H_0: \lambda = \lambda_0 = 0.6$

按 Transform→Compute Variable 顺序, 打开 Compute Variable 主对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 *卡方值对应的概率*。在 Numeric Expression 框中, 输入 “CDF.CHISQ(2*0.6*55.90,2*34)”。

则在数据编辑窗口中得到 *卡方值对应的概率* 为 0.29。

(6) 结论

由于在 $\lambda = \lambda_0 = 0.6$ 的原假设下, 出现目前统计量的值或更极端值的概率 (Sig.) 的值为 0.29, 而 $0.025 < 0.29 < 0.975$, 说明目前统计量的值落在肯定域内, 所以不拒绝 $\lambda = 0.6$ 的假设。

表 6-59 总和的计算结果

Descriptive Statistics		
	N	Sum
时间间隔	34	55.90
Valid N (listwise)	34	

第7章 非参数假设检验

在前面介绍的方法中，往往都事先假定总体服从某种特定的分布，如正态分布，然后对其均值、方差等参数作差异的显著性检验。但某个随机变量是否服从某种特定的分布是需要进行检验的。在总体分布情况不明时，用来检验总体是否来自同一个总体的假设的一类统计方法统称为非参数检验。

7.1 二项分布检验

7.1.1 二项分布检验概述

在第4章中，已经提到，如果随机变量 X 的分布如下

$$P\{X=k\} = C_n^k p^k q^{n-k} \quad (k=0,1,2,\dots,n) \\ (0 < p < 1, q = 1 - p)$$

则称 X 服从二项分布，或记为： $X \sim B(n, p)$ 。

在原假设 $H_0: p = p_0$ 时，双侧精确检验的概率为

$$2 \left(\sum_{i=0}^m \binom{N}{i} p^{*i} (1-p^*)^{N-i} \right) - \binom{N}{m} p^{*m} (1-p^*)^{N-m}$$

其中， $N = n_1 + n_2$ ， n_1 为类别1中观察值的数量， n_2 为类别2中观察值的数量。如果 $m = n_1$ ，则 $p^* = p$ ，否则， $p^* = 1 - p$ 。 p 为检验的概率，而 $m = \min(n_1, n_2)$ 。

当 $P(H_0) < \alpha$ 时，拒绝原假设。

7.1.2 二项分布检验实例分析

例 7.1 已知某校上报的达标率为 75%，经验收抽查结果为抽查的 180 人中有 127 人达标，试问该校抽查的达标结果 X 是否服从 $B(180, 0.75)$ ($\alpha = 0.05$)？

(1) 在数据编辑窗口中，建立数据文件，见 data07-01.sav。

(2) 加权处理

按 Data→Weight Cases 顺序，打开 Weight Cases 对话框，见图 2-73。选择 Weight Cases

by 选项, 在左侧变量名源框中, 选中人数变量, 并单击右移箭头将人数移入 Frequency Variable 框中, 作为加权变量。单击 OK 按钮, 完成加权设置。

(3) 按 Analyze→Nonparametric Tests→Binomial 顺序, 打开 Binomial Test 对话框, 见图 6-15。

在左侧变量名源中选择达标情况变量将其移入到 Test Variable List 下框中。

在 Test Propotion 后框中输入 0.75。

(4) 单击 Exact 按钮, 展开 Exact Tests 对话框, 见图 7-1。选择 Exact 选项。

单击 Continue 按钮返回 Binomial Test 对话框。

(5) 单击 OK 按钮运行, 则在输出窗口中, 得到计算结果, 见表 7-1。

(6) 结果与讨论

表 7-1 从左到右依次列出的是变量名 (达标情况)、事件类别 (达标、未达标)、实际观察数 (127、53)、观察概率 (0.71、0.29)、检验概率 (0.75)、原假设成立的渐近分布单侧概率 (0.100)、原假设成立的精确检验的单侧概率 (0.100)。

备注 a 指出备择假设陈述第一组样品的比例 < 0.75 。

图 7-1 Exact Tests 对话框

这说明在本程序中, 采用的是单侧检验, 原假设为第一组样品的比例大于等于 0.75。

备注 b 指出, 现有的结果是根据 Z 的渐近分布计算的。

表 7-1 二项分布假设检验结果

Binomial Test						
达标情况	Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)	Exact Sig. (1-tailed)
Group 1	达标	127	.71	.75		
Group 2	未达标	53	.29			
Total		180	1.00			

a. Alternative hypothesis states that the proportion of cases in the first group < .75.

b. Based on Z Approximation.

(7) 结论

由于原假设为第一组样品的比例大于等于 0.75 的精确检验的概率为 0.100 大于 0.05, 故不拒绝原假设, 即现有结果还不足于拒绝该校抽查的达标结果 X 服从 $B(180, 0.75)$ 。

设在本例的观察中, 记达标为 1, 未达标为 2, 则将每次的观察结果以原始记录的形式组成数据文件, 见 data07-02.sav, 则本例在 SPSS 中的操作步骤按上述的步骤在前 3 步中做如下变动:

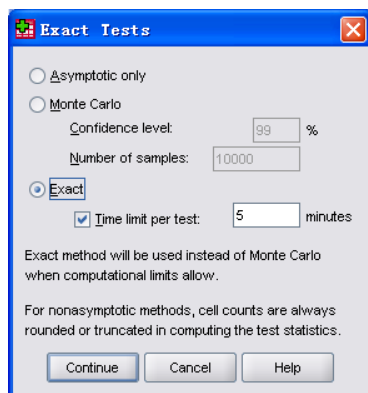
① 在数据编辑窗口中, 打开数据文件 data07-02.sav。

② 按 Analyze→Nonparametric Tests→Binomial 顺序, 打开 Binomial Test 对话框, 见图 6-15。

在左侧变量名源中选择达标情况变量将其移入到 Test Variable List 下框中。

在 Test Propotion 后框中输入 0.75。

在 Define Dichotomy (定义两分法) 中, 选择 Cut point, 并在其后框中输入 1, 即将



小于等于 1 的分成一组，大于 1 的分成另一组。

第 4 步后不变，则可得到相同的检验结果。

7.2 卡方拟合分布检验

7.2.1 对多项分布各项概率已知时卡方拟合分布检验

根据某项指标，将总体分为 r 类（对于计量资料则为 r 个连续、不重叠的区间）： A_1, A_2, \dots, A_r 。当该总体分布的参数已知时，则根据已知总体分布的概率密度函数，可以得出 A_1, A_2, \dots, A_r 各类（区间）对应的理论概率 p_{i0} 。对于一个未知分布的总体，要检验它是否同一个已知分布的总体属同分布，可以通过从该未知分布的总体中随机抽取一个容量为 n 的样本，则当未知分布的总体与已知分布的总体同分布时，由样本所确定的 A_1, A_2, \dots, A_r 各类所占的比例 p_i 与 p_{i0} 之间应是一致的。因此，分布的一致性检验，等价于作如下的假设检验：

H_0 ：类 A_i 所占的比例为 $p_i = p_{i0}$ ($i=1, 2, \dots, r$)。

在原假设 H_0 成立时，每一类 i 中观察到的样本中实际出现的频数 O_i ($O_i = n_i$) 与理论上的期望频数 E_i ($E_i = np_{i0}$) 之间应相当接近，故著名统计学家皮尔逊提出用统计量

$$\chi^2 = \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i}$$

作为衡量实际频数与理论期望频数的偏差的综合指标，并证明了，在原假设 H_0 成立时，上述统计量的渐近分布为 $\chi^2(r-1)$ 。在给定显著性水平 α 下，当 $\chi^2 \geq \chi_{1-\alpha}^2(r-1)$ ，拒绝原假设。

在 SPSS 下拉式菜单 Analyze 的 Nonparametric Tests 过程的 Chi-Square 程序中，可以用上法来进行拟合分布检验。它提供了以下三种检验方法：

第一种为 Asymptotic only，即渐近分布，就是上面提到的方法。它的使用条件为大样本，期望频数大于 5。

第二种为 Monte Carlo 法，它是一种计算确切概率值的方法，在样本量较大，但又有期望频数小于 5 的情况出现时使用。

第三种为 Exact 法，即精确检验法，当期望频数小于 5 的情况出现时使用。它能精确地计算原假设成立的概率值，但耗时长，易受计算机内存限制，当样本容量小于 30 时，计算较快。

例 7.2 19 世纪伟大的英国生物学家孟德尔做过一个豌豆试验，他按颜色与形状把豌豆分为四类：黄而圆的，青而圆的，黄而有角的，青而有角的。根据遗传学显性

因子和隐性因子的遗传特性，孟德尔假设子系从父系（母系）接受显性因子“黄色”和隐性因子“青色”（或显性因子“圆”和隐性因子“有角”）的概率都等于 1/2。由此，他指出这四类的豌豆个数之比为 9：3：3：1。在随机抽取的 556 颗豌豆中，他观察到上述四类的豌豆数分别为 315、108、101、32。试问观察结果是否拒绝孟德尔的假设？

【题析】本例所要做的原假设为 H_0 ：四类豌豆出现的概率分别为 9/16、3/16、3/16、1/16。

在 SPSS 中的具体做法如下：

- (1) 在 SPSS 数据编辑窗口中，建立数据文件，见 data07-03.sav。
- (2) 加权处理

按 Data→Weight Cases 顺序打开 Weight Cases 对话框，见图 2-73，选择 Weight Cases By 选项，并将频数变量移入 Frequency Variable 下框中，单击 OK 按钮返回数据编辑窗，完成加权处理。

(3) 按 Analyze→Nonparametric Tests→Chi-Square 顺序打开 Chi-Square 对话框，见图 3-17。

将类别移入到 Test Variable List 下框中。选择 Values 选项，意指各类别的期望值（概率）是不一样的。由于程序会对各类别输入的值自动地进行归一化处理，所以，本例可以按 9：3：3：1 的比例依次输入即可。做法如下：

在 Values 选项后框中输入对应于第一类比例值 9，单击 Add 按钮，将其增加到下框中，再在 Values 后框中输入对应于第二类比例值 3，单击 Add 按钮，将其增加到下框中。直到将第四类比例值 1 添加完为止。见图 3-17。

(4) 单击 Exact 按钮，弹出如图 7-1 所示的精确检验选项卡，由于本例为大样本，为节省计算时间，故采用系统默认选项 Asymptotic only（渐近分布）。

单击 Continue 按钮，返回 Chi-Square 对话框，见图 3-17。

(5) 单击 OK 按钮，在输出窗口中，得到本次选择的计算结果，见表 7-2 和表 7-3。

(6) 结果解释

表 7-2 频数分布

类别			
	Observed N	Expected N	Residual
黄而圆的	315	312.8	2.2
青而圆的	108	104.2	3.8
黄而有角的	101	104.2	-3.2
青而有角的	32	34.8	-2.8
Total	556		

表 7-3 检验结果

Test Statistics	
	类别
Chi-Square	.470 ^a
df	3
Asymp. Sig.	.925
a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 34.8.	

在表 7-2 中,第一列列出了分类变量的四种类别,第二列给出了各类别的实际观察数,第三列给出的是各类别期望的频数,第四列给出的是实际观察值与期望频数之间的差值,称残差。

在表 7-3 中,显示了卡方检验的结果,卡方统计量值为 0.470,自由度 df 为 3,在观察值与期望频数无差异的原假设下,出现目前统计量的值或者更极端值的概率(Asymp.Sig.)为 0.925 大于 0.05,表下面的备注说明,单元格中最小期望频数为 34.8,没有出现期望频数小于 5 的情况,说明使用系统默认选项 Asymptotic only(渐近分布)来检验是合适的。所以不拒绝原假设,也就是观察结果不拒绝孟德尔的假设。

例 7.3 某电话交换台,在 100 分钟内记录了每分钟被呼叫的次数 x_i ,见表 7-4,其中 n_i 为出现呼叫次数 x_i 的频数。可否认为每分钟被呼叫的次数服从 $\lambda = 4.5$ 的泊松分布?

表 7-4 100 分钟内每分钟被呼叫的次数 x_i 的记录表

x_i	0	1	2	3	4	5	6	7	8	≥ 9
n_i	1	7	12	18	17	20	13	6	3	3

本例所作的原假设为: $H_0: X \sim P(4.5)$ 。

在 SPSS 中的解题步骤如下:

(1) 在 SPSS 数据编辑窗口中,建立数据文件,见 data07-04.sav。

(2) 加权处理

按 Data→Weight Cases 顺序打开 Weight Cases 对话框,见图 2-73,选择 Weight Cases By 选项,并将频数变量移入 Frequency Variable 下框中,单击 OK 按钮返回数据编辑窗,完成加权处理。

(3) 计算理论频数

按 Transform→Compute Variable 顺序,打开 Compute Variable 主对话框(见图 2-53)。在 Target Variable 框中,输入目标变量名为理论概率。在 Numeric Expression 框中,输入“PDF.POISSON(呼叫次数,4.5)”。

(4) 单击 OK 按钮运行,则在数据编辑窗口工作的数据文件中,出现理论概率的新变量及其值。见图 7-2。

(5) 修正最后一组理论概率值

用 1-CDF.POISSON(8,4.5)可计算得到呼叫次数大于等于 9 次时的理论概率为 0.04。

呼叫次数	频数	理论概率
0	1	0.01
1	7	0.05
2	12	0.11
3	18	0.17
4	17	0.19
5	20	0.17
6	13	0.13
7	6	0.08
8	3	0.05
9	3	0.02

图 7-2 理论频数计算结果

(6) 进行卡方检验

按 Analyze→Nonparametric Tests→Chi-Square 顺序打开 Chi-Square 对话框, 见图 3-17。

将 **呼叫次数** 移入到 Test Variable List 下框中。选择 Values 选项, 意指各类别的期望值 (概率) 是不一样的。在 Values 选项后框中输入对应于呼叫次数 0 的理论概率值 0.01, 单击 Add 按钮, 将其增加到下框中, 再在 Values 后框中输入对应于呼叫次数 1 的理论概率值 0.05, 单击 Add 按钮, 将其增加到下框中。直到将呼叫次数 9 的理论概率值 0.04 添加完为止。

(7) 单击 Exact 按钮, 弹出如图 7-1 所示的精确检验对话框, 由于本例中会出现理论频数小于 5 的情况, 故采用 Exact 选项, 其他采用系统默认值。

单击 Continue 按钮, 返回 Chi-Square 对话框, 见图 3-17。

(8) 单击 OK 按钮, 在输出窗口中, 得到本次选择的计算结果, 见表 7-5 和表 7-6。

表 7-5 频数分布

呼叫次数			
	Observed N	Expected N	Residual
0次	1	1.0	.0
1次	7	5.0	2.0
2次	12	11.0	1.0
3次	18	17.0	1.0
4次	17	19.0	-2.0
5次	20	17.0	3.0
6次	13	13.0	.0
7次	6	8.0	-2.0
8次	3	5.0	-2.0
>=9次	3	4.0	-1.0
Total	100		

表 7-6 检验结果

Test Statistics	
	呼叫次数
Chi-Square	3.240 ^a
df	9
Asymp. Sig.	.954
Exact Sig.	.957
Point Probability	.000

a. 2 cells (20.0%) have expected frequencies less than 5. The minimum expected cell frequency is 1.0.

(9) 结果解释

在表 7-5 中, 第一列列出了每分钟被呼叫次数的 10 种情况, 第二列给出了每种情况的实际观察数, 第三列给出的是各种情况的期望频数, 第四列给出的是实际观察值与期望频数之间的差值, 称残差。

在表 7-6 中, 显示了卡方检验的结果, 卡方统计量值为 3.240, 自由度 df 为 9, 原假设成立的概率 Exact Sig. 为 0.957 大于 0.05, 表下面的备注说明, 单元格中最小期望频数为 1.0, 有 20% 的期望频数小于 5, 说明不能使用系统默认选项 Asymptotic only (渐近分布) 来检验。所以不拒绝原假设, 也就是认为每分钟被呼叫的次数服从 $\lambda = 4.5$ 的泊松分布。

7.2.2 对多项分布各项概率未知时的卡方拟合分布检验

当总体分布的类型已知, 而总体分布的参数未知时, 将总体分为 r 类 (对于计量资料则为 r 个连续、不重叠的区间): A_1, A_2, \dots, A_r 时, 各类的 p_i 值是未知的, 此时, 需要对 7.2.1 中的方法进行修正, 即需要用从样本中计算得到的总体参数的最大似然估计的估

计量来代替总体的未知参数, 此时, 理论上的期望频数 $E_i = n\hat{p}_{i0}$, 再用统计量

$$\chi^2 = \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i}$$

来计算 χ^2 值, 在原假设 $H_0: p_i = \hat{p}_{i0}$ 成立时, $\chi^2 \sim \chi^2(r - m - 1)$ 。其中 m 为总体参数的个数。

在实际应用时, 还有一个问题需要注意, 那就是对总体, 也即检验指标 x 值域的划分方式。一般地, 划分可按自然形式并将临近的值分在一起 (如按频数分布表法划分)。但应当使每一组的期望频数 E_i 都保持在一定的数目之上。如果有的 E_i 太小, 那么在计算 χ^2 值时, 相应的 $\frac{(O_i - E_i)^2}{E_i}$ 波动就会过大。 E_i 至少应为多大, 不同的学者有不同的看法, 保守一点的认为应大于 5, 也有的认为大于等于 3 即可。我们建议当 $E_i < 5$ 时, 临近组合并。

例 7.4 某农科站为了考察某种大麦穗长的分布情况, 在一块试验地里随机抽取了 100 个麦穗, 其整理后得到的频数分布数据见 data07-05.sav。试检验麦穗长度是否服从正态分布 ($\alpha = 0.05$) ?

在 SPSS 中解题步骤如下:

1. 计算样本均值和标准差

- (1) 在 SPSS 数据编辑窗口打开 data07-05.sav。
- (2) 用 频数 作加权变量对数据文件进行加权处理。
- (3) 对频数分布数据资料计算样本均值和标准差。

在 SPSS 中有很多种做法都可求得频数分布资料的均值和标准差的, 以下只是众多做法中的一种。

按 Analyze → Reports → Case Summaries 打开 Case Summaries 的对话框, 见图 2-51。在左侧的源变量表中选择 组上 变量将其送入 Variables 框中。关闭 Display Cases 选择项, 单击 Statistics 按钮, 展开 Statistics 选择项框, 见图 2-52。将统计量 Mean (样本均值)、Standard Deviation (标准差) 移入右框中。单击 Continue 按钮, 返回图 2-51。单击 OK 按钮运行, 在输出窗中得到表 7-7 的输出结果。

表 7-7 总体参数估计量
Case Summaries

组上	
Mean	Std. Deviation
6.1090	.60371

一般而言, 在频数分布资料中, 其均值可用下式求得

$$\bar{x} = \frac{\sum_{i=1}^k f_i M_i}{\sum_{i=1}^k f_i}$$

式中 f_i 为第 i 组的组内频数, M_i 为第 i 组的组中值。在等距离分组的频数分布资料中, 组中值可以通过相邻组的组上限之和除以 2 (或相邻组的组下限之和除以 2) 来获取。我们知道, 当一个数组的每个元素都加上一个常量后, 新数组的均值比原数组的均值要大一个加上去的常量, 但两个数组的标准差值不受影响, 仍然相等。因此, 在由小到大排列的频数分布表中, 通过组上限求得的所谓均值, 它实际上要比由组中值求得的均值大一个常数值, 这个常数值为任一组的组上限减去该组的组中值, 本例中, 按第二组为例, 它大 $4.6 - (4.6+4.3)/2=0.15$, 所以, 本例中的标准的均值应为 $6.109-0.15=5.959$, 而标准差仍为 0.60371。

2. 计算各组的累计概率

按 Transform→Compute Variable 顺序, 打开 Compute Variable 主对话框 (见图 2-53)。在 Target Variable 框中, 输入目标变量名为 *组内累计概率*。在 Numeric Expression 框中, 输入 “CDF.NORMAL(组上限,5.959,0.60371)”。

单击 OK 按钮运行, 则在数据编辑窗口工作的数据文件中, 出现 *组内累计概率* 的新变量及其值。并按 File→Save as 顺序, 打开 Save Data As 对话框, 将其另存为 data7-06.sav。

注: 现在工作的数据文件为 data7-06.sav。

3. 计算各组内概率

按 Transform→Creat time series 顺序, 打开 Creat time series 对话框, 见图 3-15。

在左侧源变量表中, 选择 *组内累计概率* 并将其移入到 New Variable[s] 下框中, 在 Name and Function 下, Name 后框中输入 *各组内概率*, 其他不做选择, 即采用系统默认的 1 阶差分。

单击 OK 按钮运行, 则在当前工作的数据文件中, 出现 *各组内概率* 的新变量及其值。见图 7-3。

组别	组上限	频数	组内累计概率	各组内概率
1	4.30	1.00	0.0030	.
2	4.60	1.00	0.0122	0.0092
3	4.90	2.00	0.0397	0.0275
4	5.20	5.00	0.1043	0.0646
5	5.50	12.00	0.2235	0.1192
6	5.80	15.00	0.3961	0.1726
7	6.10	28.00	0.5923	0.1962
8	6.40	13.00	0.7675	0.1751
9	6.70	10.00	0.8902	0.1227
10	7.00	10.00	0.9577	0.0675
11	7.30	2.00	0.9868	0.0292
12	7.60	1.00	0.9967	0.0099

图 7-3 计算结果

组上限	频数	组内累计概率	各组内概率
5.20	9.00	0.1043	0.1043
5.50	12.00	0.2235	0.1192
5.80	15.00	0.3961	0.1726
6.10	28.00	0.5923	0.1962
6.40	13.00	0.7675	0.1751
6.70	10.00	0.8902	0.1227
7.60	13.00	0.9967	0.1066

图 7-4 合并后的结果

第 1 组的组内概率等于第 1 组的组内累计概率。

4. 为满足卡方检验的基本要求, 对组内频数不足 5 的人工进行合并, 处理后的结果见图 7-4, 并将其存放在 data07-07.sav 中。

5. 进行卡方检验

在 SPSS 数据编辑窗口中, 打开数据文件 data07-07.sav。按 Analyze→Nonparametric Tests→Chi-Square 顺序打开 Chi-Square 对话框, 见图 3-17。

将组上限移入到 Test Variable List 下框中。选择 Values 选项, 意指各类别的期望值(概率)是不一样的。在 Values 选项后框中输入对应于组上限 5.20 组的理论概率值 0.1043, 单击 Add 按钮, 将其增加到下框中, 再在 Values 后框中输入对应于组上限 5.50 组的理论概率值 0.1192, 单击 Add 按钮, 将其增加到下框中。直到组上限 7.60 组的理论概率值 0.1066 添加完为止。

6. 单击 Exact 按钮, 弹出如图 7-1 所示的精确检验对话框, 由于本例中可能会出现理论频数小于 5 的情况, 故选用 Monte Carlo 选项, 其他采用系统默认值。

单击 Continue 按钮, 返回 Chi-Square 对话框, 见图 3-17。

7. 单击 OK 按钮, 在输出窗口中, 得到本次选择的计算结果, 见表 7-8 和表 7-9。

表 7-8 频数分布

组上限	Observed N	Expected N	Residual
5.2	9	10.5	-1.5
5.5	12	12.0	.0
5.8	15	17.3	-2.3
6.1	28	19.7	8.3
6.4	13	17.6	-4.6
6.7	10	12.3	-2.3
7.6	13	10.7	2.3
Total	100		

表 7-9 检验结果

Test Statistics		组上限
Chi-Square		6.146 ^a
df		6
Asymp. Sig.		.407
Monte Carlo Sig.	Sig.	.403 ^a
99% Confidence Interval		
Lower Bound		.391
Upper Bound		.416

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 10.5.

b. Based on 10000 sampled tables with starting seed 2000000.

8. 结果与讨论

在表 7-8 中, 第一列列出了所分的 12 个组的各组组上限, 第二列给出了每组的实际观察数, 第三列给出的是各组的期望频数, 第四列给出的是实际观察值与期望频数之间的差值, 称残差。

在表 7-9 中, 显示了卡方检验的结果, 卡方统计量值为 6.146, 自由度 df 为 $7-2-1=4$, 这与表中的 6 是不一样的, 因为本例中的正态总体参数有两个, 是未知的, 因而它是根据样本统计量估计的。所以, 自由度同表中给出的是有区别的。因而表中原假设成立的概率也是不能用的。

它可将计算得到的卡方值和自由度代入到函数 $CDF.CHISQ(6.146,4)$ 中计算得到在 6.146 处的累计概率, 本例为 0.8115, 由于 $0.025 < 0.8115 < 0.975$, 故 $P(H_0) > 0.05$ 。

9. 结论

不能拒绝麦穗长度服从正态分布的假设。

由上可知, 在总体分布情况不明时, 可以根据以往的经验或实际的观测数据的分布

情况, 推测总体可能服从某种分布函数 $F(x)$, 利用这些样本数据来具体检验该总体分布函数是否真的就是 $F(x)$ 。卡方检验 (Chi-square test) 就是这样一种用来检验给定的概率值下数据来自同一总体的无效假设的方法。

虽然, 卡方检验主要用于分布的一致性检验, 但也可推广应用于两个或多个总体间率的差异的显著性检验。例如, 总体分布的正态性检验, 一种新药对某种疾病的治愈率和原有药物治愈率的比较, 不同教学法下学生的达标率间是否有差异等, 都可用卡方检验的方法来加以分析。

7.3 序列随机性的游程检验

7.3.1 游程检验概述

7.3.1.1 游程

一个游程就是某个两分变量序列中位于一种符号 (或一个值) 之前或之后的另一种符号 (或另一个值) 持续的最大主序列, 或者说, 一个游程是指某序列中同类元素的一个持续的最大主集。

例如, 在做 30 次抛掷一枚硬币的试验中, 得到如下的观察结果:

000011100000110000011111100010,

其中数字 0 表示出现正面在上的结果, 数字 1 表示出现反面在上的结果。

根据游程的定义可知, 在上面 30 次试验结果中, 连续出现的 0 和连续出现的 1, 均被称为游程, 因此, 它共有 7 个游程 ($R=7$), 其中 4 个为 0 的游程和 3 个为 1 的游程。

根据做抛掷均匀硬币试验的经验可知, 出现多个 0 或多个 1 连续地连在一起的情况是不多见的, 而 0 和 1 交替频繁地出现也是不太可能的。因此, 一个有太多或太少游程的样本暗示着该样本不是随机的。

7.3.1.2 游程检验 (Runs Test)

它是根据游程数来检验变量的两个值 (或符号) 出现的顺序是否是随机的一种非参数检验方法。

其基本做法如下:

1. 首先定义一个分界点, 即指定一个特定的数或计算得到的统计量的值来两分数据。可能的统计量为: 样本均值、中位数、众数。

2. 用给定的一个分界点, 计算数据文件序列与其差值: $D_i = X_i - \text{分界点}$ 。如果 $D_i \geq 0$, 差值为正, 记为 1, 否则为负, 记为 0。

3. 计算 1 出现的次数 n_p 、0 出现的次数 n_a 和符号改变 (即 $D_i \geq 0$ 而 $D_{i+1} < 0$, 或 $D_i < 0$ 而 $D_{i+1} \geq 0$) 的次数, 则总的样本含量为 $N = n_p + n_n$, 游程数量 R 为符号变化的

次数加 1。

4. 假设检验

游程检验中所要作的原假设为 H_0 ：两分变量有随机性。

在原假设成立的前提下，当样本容量很大，游程数量 R 的抽样分布近似服从 $N(\mu_r, \sigma_r)$ 。因此，统计量

$$Z = \frac{R - \mu_r}{\sigma_r} \sim N(0,1)$$

$$\text{其中 } \mu_r = \frac{2n_p n_a}{n_p + n_a} + 1, \quad \sigma_r = \sqrt{\frac{2n_p n_a (2n_p n_a - n_a - n_p)}{(n_p + n_a)^2 (n_p + n_a - 1)}}$$

在 $n < 50$ 时，对上述计算公式进行如下修正

$$Z_c = \begin{cases} (R - \mu_r + 0.5)/\sigma_r & \text{如果 } R - \mu_r \leq 0.5 \\ (R - \mu_r - 0.5)/\sigma_r & \text{如果 } R - \mu_r \geq 0.5 \\ 0 & \text{如果 } |R - \mu_r| < 0.5 \end{cases}$$

当 $P(H_0) < \alpha$ 时，拒绝原假设 H_0 。

游程数 R 点概率的计算：

当 R 为偶数时，设 $R = 2k$ ， $k = 1, 2, \dots, n_p$ ，则游程数 R 点概率为

$$P([R = 2k]) = \frac{2 \binom{n_n - 1}{k - 1} \binom{n_p - 1}{k - 1}}{\binom{N}{n_n}}$$

当 R 为奇数时，设 $R = 2k + 1$ ，当 $n_n > n_p$ ， $k = 1, 2, \dots, n_p$ ，当 $n_n = n_p$ ， $k = 1, 2, \dots, n_p - 1$ ，则游程数 R 点概率为

$$P([R = 2k + 1]) = \frac{\binom{n_n - 1}{k - 1} \binom{n_p - 1}{k} + \binom{n_n - 1}{k} \binom{n_p - 1}{k - 1}}{\binom{N}{n_p}}$$

5. 游程检验的应用

游程检验可用来检验样本的随机性，这对于统计推断是很重要的。游程检验也可用来检验任何序列的随机性，而不管这个序列是怎样产生的。此外还可用来判断两个总体的分布是否相同，从而检验出它们的位置中心有无显著差异。

7.3.1.3 游程检验实例分析

1. 序列的随机性检验

例 7.5 统计学课上，老师布置了让学生做抛掷硬币 60 次的实验，要求用数字 0 记录出现正面在上的结果，用数字 1 记录出现反面在上的结果，在学生所交的作业中，有一位学生的实验结果为：000011100000110000011111100010000011100000110000011111100010，试问该生真的做了实验了吗？

【题析】 如果学生真的做了实验，则硬币出现正面在上和反面在上的几率是几乎相同的，所以，从游程数上看，它不会太多也不会太少，正面在上的次数和反面在上的次数大体相同，因此，实验得到的序列应是随机的。反之，则很可能是学生未做实验的结果。

在 SPSS 中，本例的解题步骤为：

(1) 在 SPSS 中，将题中数据建成数据文件，见 data07-08。

(2) 按 Analyze→Nonparametric Tests→Runs 顺序展开 Runs Test 对话框，见图 7-5。

(3) 选择实验结果变量送入 Test Variable 对话框。

(4) 本例中由于 1 出现 24 次，0 出现 36 次，中位数和众数都为 0，故不能选用 Mode（众数）和 Median（中位数）作为划分两类的分界点。本例选用 Mean（平均数）。如果选择 Custom，则分割点值应大于 0 且小于 1。

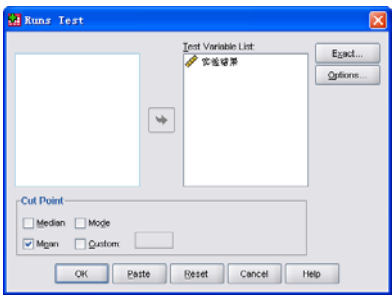


图 7-5 游程检验对话框

(5) 单击 Exact 按钮，弹出如图 7-1 所示的精确检验对话框，选择 Exact 选项，其他用系统默认值。

单击 Continue 按钮，返回 Runs Test 对话框，见图 7-5。

(6) 单击 OK 按钮运行，在输出窗口中出现计算结果，见表 7-10。

(7) 结果解释

Runs Test 表是以 Mean 平均数作为分界点的结果，计算结果平均数是 0.40。

Runs Test 表的第二行起依次为：检验值（即分界点值，本例为样本数据的均值 0.40）、小于检验值的样品数为 36、大于等于检验值的样品数为 24、总样品数 60、游程数量 17、Z 值为-3.475，在正反面出现的机会是随机的原假设下，出现目前统计量的值或者更极端值的双侧检验概率为 0.001，精确检验的双侧检验概率值为

表 7-10 游程检验结果

Runs Test	
	实验结果
Test Value ^a	.40
Cases < Test Value	36
Cases ≥ Test Value	24
Total Cases	60
Number of Runs	17
Z	-3.475
Asymp. Sig. (2-tailed)	.001
Exact Sig. (2-tailed)	.001
Point Probability	.000

a. Mean

0.001, 游程数 17 点的概率为 0.000。因 $P=0.001<0.05$, 故拒绝原假设, 即该学生做的掷硬币试验不是随机的, 因此认为该生没有真的做了实验, 实验结果是编造的。

2. 样本数据的随机性检验

例 7.6 对某一个总体进行了观察, 所得到的数据按观察顺序的先后排列如下:

18, 34, 5, 16, 58, 37, 18, 14, 21, 48, 43, 22, 53, 36, 38, 9, 15, 63, 56, 64, 26, 30, 33, 50, 3, 60, 41。

试用游程检验法检验假设“这组观察值是随机的”。

【题析】 在如本例的连续型计量资料的具体实际问题中, 并不是所有的数据对都是以 0 或 1 的二元形式来表现的。它可先找出中位数, 然后所有的原始数据与中位数来比较, 大于中位数的计为 1, 小于中位数的计为 0, 这样可把计量资料变成一组 0、1 序列。就可按两分变量的随机性方法来做检验了。在 SPSS 中, 在单样本情况下, 这种将原始观察值转换成 0 或 1 的过程由程序自动完成, 不必手工来做。故在 SPSS 中的操作步骤如下:

- (1) 在 SPSS 中, 将题中数据建成数据文件, 见 data07-09。
- (2) 按 Analyze→Nonparametric Tests→Runs 顺序展开 Runs Test 对话框, 见图 7-5。
- (3) 选择 *观察值* 变量送入 Test Variable 对话框。
- (4) 在 Cut Point 中选用 Median (中位数) 作为划分两类的分界点。

(5) 单击 Exact 按钮, 弹出如图 7-1 所示的精确检验选项卡, 选择 Exact 选项, 其他用系统默认值。

单击 Continue 按钮, 返回 Runs Test 对话框, 见图 7-5。

(6) 单击 OK 按钮运行, 在输出窗口中出现计算结果, 见表 7-11。

表 7-11 游程检验结果

(7) 结果解释

Runs Test 表是以 Median (中位数) 作为分界点的结果, 计算结果中位数是 34.00。

Runs Test 表的第二行起依次为: 检验值 (即分界点值, 本例为样本数据的中位数 34.00)、小于检验值的样品数为 13、大于等于检验值的样品数为 14、总样品数 27、游程数量 14、Z 值为 0.000, 在样本数据是随机的原假设下, 出现目前统计量的值或者更极端值的双侧检验概率为 1.000, 精确检验的双侧检验概率值为 1.000, 游程数 17 点的概率为 0.158。因 $P=1.000>0.05$, 故不拒绝原假设, 即认为这组观察值是随机的。

3. 两个样本的同分布检验

例 7.7 从两个总体中, 相互独立地各随机抽取一个样本, 得到表 7-12 所示的观察值, 试用游程检验法检验这两个总体是否同分布?

Runs Test	
	观察值
Test Value ^a	34.00
Cases < Test Value	13
Cases ≥ Test Value	14
Total Cases	27
Number of Runs	14
Z	.000
Asymp. Sig. (2-tailed)	1.000
Exact Sig. (2-tailed)	1.000
Point Probability	.158
a. Median	

表 7-12 两组样本数据

X	11.2	13.0	13.2	10.0	14.2	15.0	14.1	16.1	17.1	15.8	14.0	13.6	9.9
Y	15.1	16.2	17.2	18.0	16.9	14.7	14.5	16.8	14.7	13.7			

【题析】 本例所要作的原假设是 H_0 ：总体 X 和 Y 的分布相同。在原假设为真时，我们可以对其中一组数据做上标记后，进行两组数据合并后排序，然后分别用 0 和 1 标识两组数据，得到 0、1 序列。如果样本来自的两总体的分布形态存在较大差距，则计算出的游程数会相对比较小。如果游程数比较大，则应是由于两样本数据充分混合的结果，则它们的分布应该不存在显著差异。

在 SPSS 中解题时，可用下述步骤进行：

(1) 在 SPSS 中，将题中数据建成数据文件，见 data07-10。

(2) 对观察值排序

按 File→Data→Sort Cases 顺序打开 Sort Cases 对话框，见图 2-26。将观察值变量拖曳到 Sort by 框中，采用系统默认的 Ascending（升序）进行排序，单击 OK 按钮，实现对原数据文件中的样品进行按观察值的大小由小到大进行重排。

在数据编辑窗口可以看到，组别变量的值也随之进行了重新排列，它便是我们想要得到的对应观察值混合后的有序两分序列。

(3) 按 Analyze→Nonparametric Tests→Runs 顺序展开 Runs Test 对话框，见图 7-5。

(4) 选择组别变量送入 Test Variable 对话框。

(5) 在 Cut Point 中选用 Mean（均值）作为划分两类的分界点。

(6) 单击 Exact 按钮，弹出如图 7-1 所示的精确检验对话框，选择 Exact 选项，其他用系统默认值。

单击 Continue 按钮，返回 Runs Test 对话框，见图 7-5。

(7) 单击 OK 按钮运行，在输出窗口中出现计算结果，见表 7-13。

(8) 结果解释

Runs Test 表是以 Mean（均值）作为分界点的结果，计算结果均值是 0.43。

Runs Test 表的第二行起依次为：检验值（即分界点值，本例为样本数据的均值 0.43）、小于检验值的样品数为 13、大于等于检验值的样品数为 10、总样品数 23、游程数量 10、Z 值为 -0.784，在两个样本是同分布的原假设下，出现目前统计量的值或者更极端值的双侧检验概率为 0.433，精确检验的双侧检验概率值为 0.388，游程数 10 点的概率为 0.109。因 $P=0.383>0.05$ ，故不拒绝原假设，即认为这两个样本所在

表 7-13 游程检验结果

Runs Test	
	组别
Test Value ^a	.43
Cases < Test Value	13
Cases ≥ Test Value	10
Total Cases	23
Number of Runs	10
Z	-.784
Asymp. Sig. (2-tailed)	.433
Exact Sig. (2-tailed)	.388
Point Probability	.109

a. Mean

的总体是同分布的。

7.4 柯尔莫哥洛夫-斯米诺夫检验

7.4.1 柯尔莫哥洛夫-斯米诺夫检验基本概述

假定总体 X 的分布函数 $F(x)$ 连续但未知, 要在给定的显著性水平 α 下检验假设

$$H_0: F(x) = F_0(x); H_1: F(x) \neq F_0(x)$$

虽然, 这个问题可用 χ^2 拟合分布检验来完成, 但有不足之处。这是由于 χ^2 拟合优度检验是比较样本频率与总体概率之间的差异, 这势必会受到区间划分的局限。这会导致原假设不真时, 而在某种区间的划分下, 把不真的原假设接受下来。

柯尔莫哥洛夫-斯米诺夫检验克服了 χ^2 拟合分布检验的缺陷, 它不是在划分区间上考虑未知分布与已知分布间的偏差, 而是在每一个点上考虑两者之间的偏差, 它需要假定总体分布是连续的。这使得它的应用有局限性, 这是它唯一的不足。

在 SPSS 的非参数假设检验中, 一个样本的柯尔莫哥洛夫-斯米诺夫检验 (One-Sample Kolmogorov-Smirnov Test) 可用来检验样本来自正态 (Normal)、均匀 (Uniform) 或泊松分布 (Poisson) 总体的假设。

Kolmogorov-Smirnov 双侧检验的原假设 H_0 为: 对所有的 x 值 $F(x)=F(x_0)$ 成立, 备择假设为: 至少有一个 x 值使 $F(x) \neq F(x_0)$ 成立。

设 $S(x)$ 表示一组数据的经验分布。定义一组随机样本 x_1, x_2, \dots, x_n 的经验分布函数为阶梯函数

$$S(x) = \frac{x_i \leq x \text{ 的个数}}{n}$$

它是小于 x 的值的比列, 是总体分布 $F(x)$ 的一个估计。检验统计量为

$$D = \sup_x |S(x) - F_0(x)|$$

D 的分布对一切连续分布 $F(x_0)$ 在原假设下是一样的, 所以它与分布无关。在实际运算中, 由于 $S(x)$ 是阶梯函数, 只取离散值, 所以考虑到跳跃问题, 如果有 n 个观察值, 则可用下面的统计量来代替上面的 D

$$D = \max_{1 \leq i \leq n} \{ \max(|S(x_i) - F_0(x_i)|, |S(x_{i-1}) - F_0(x_i)|) \}$$

当 $n \rightarrow \infty$ 时, 大样本的渐近公式为

$$P(\sqrt{n}D_n < x) \rightarrow K(x)$$

其分布函数的表达式为

$$K(x) = \begin{cases} 0 & x < 0 \\ \sum_{j=-\infty}^{\infty} (-1)^j \exp(-2j^2 x^2) & x > 0 \end{cases}$$

当 $P(H_0) < \alpha$ 时, 拒绝原假设。

需要注意的是, 用本程序进行检验时, 由于已知分布的参数是要用样本中计算得到的统计量来估计的, 所以用渐近分布进行检验时, 需要大样本, 即 $n \geq 100$ 。当样本含量较少时, 应选用精确检验或 MONTE CARLO 检验。

7.4.2 实例分析

由于在例 3.10 和例 3.13 中已介绍过用柯尔莫哥洛夫-斯米诺夫方法检验样本数据资料的正态分布和泊松分布, 因此, 在此对这两个分布的检验不再赘述, 而是再介绍用它对指数分布和均匀分布进行检验的两个实例。

1. 指数分布检验实例

例 7.8 对某设备进行寿命试验, 测得 10 次无故障工作时间为: 1380, 920, 2350, 420, 500, 2100, 2300, 1760, 1650, 1510, 问此设备的无故障工作时间是否服从指数分布?

在 SPSS 中, 解题步骤如下:

(1) 在 SPSS 数据编辑窗口中, 建立数据文件。具体内容参见 data07-11.sav。

(2) 按 Analyze→Nonparametric Tests→1-Sample K-S 顺序, 打开 One-Sample Kolmogorov-Smirnov Test 对话框, 见图 3-7。

在左边变量名源框中, 选择无故障工作时间变量将其移入到 Test Variable List 框中, 在 Test Distribution 选项中只选择 Exponential (指数分布) 选项。

(3) 单击 Exact 按钮, 展开 Exact Tests 对话框, 见图 7-1。选择 Exact 选项。

单击 Continue 按钮返回 1-Sample K-S Test 对话框。

(4) 单击 OK 按钮执行, 在输出窗口中, 得到表 7-14 所示的输出结果。

第一行样本量为 10, 第二行指数参数均

表 7-14 无故障工作时间的指数分布精确检验

One-Sample Kolmogorov-Smirnov Test		无故障时间
N		10
Exponential parameter ^a	Mean	1489.0000
Most Extreme Differences	Absolute	.304
	Positive	.206
	Negative	-.304
Kolmogorov-Smirnov Z		.962
Asymp. Sig. (2-tailed)		.313
Exact Sig. (2-tailed)		.256
Point Probability		.000

a. Test Distribution is Exponential.

值为 1489.0, 第三行最大极端差值的绝对值为 0.304、正值为 0.206、负值为 -0.304, 第四行 Kolmogorov-Smirnov 的 Z 值为 0.962, 第五行渐近分布双侧检验的概率为 0.313, 第六行精确检验的双侧检验的概率为 0.256, 点概率为 0.000。

本例的样本只有 10 很小, 故不能采用渐近分布双侧检验的概率作为原假设成立的概率。只能使用精确检验的双侧检验的概率。

由于精确检验的双侧检验的概率为 0.256 大于 0.05, 所以不拒绝无故障工作时间服从指数分布的原假设。

2. 均匀分布检验实例

例 7.9 据经验知, 机床发生故障的频率服从均匀分布, 某车间在一周内统计所有车床发生故障的频数资料如表 7-15, 试用柯尔莫哥洛夫-斯米诺夫检验检验故障频数是否服从均匀分布?

表 7-15 某车间在一周内所有车床发生故障的频数资料

星期	一	二	三	四	五	六
故障频数	7	8	3	9	16	17

在 SPSS 中, 解题步骤如下:

(1) 在 SPSS 数据编辑窗口中, 建立数据文件。具体内容参见 data07-12.sav。

(2) 用故障频数作为权重变量进行加权处理。

(3) 按 Analyze→Nonparametric Tests→1-Sample K-S 顺序, 打开 One-Sample Kolmogorov-Smirnov Test 对话框, 见图 3-7。

在左边变量名源框中, 选择故障频数变量将其移入到 Test Variable List 框中, 在 Test Distribution 选项中只选择 Uniform (均匀分布) 选项。

(4) 单击 Exact 按钮, 展开 Exact Tests 选项卡, 见图 7-1。选择 Exact 选项。

单击 Continue 按钮返回 1-Sample K-S Test 对话框。

(5) 单击 OK 按钮执行, 在输出窗口中, 得到表 7-16 所示的输出结果。

表中各行的解释见例 7.8 中的表 7-14。

由于精确检验的双侧检验的概率为 0.000 小于 0.05, 所以拒绝机床发生故障的频率服从均匀分布的原假设。

表 7-16 故障频率均匀分布精确检验

One-Sample Kolmogorov-Smirnov Test		时间
N		60
Uniform Parameters*	Minimum	1
	Maximum	6
Most Extreme Differences	Absolute	.350
	Positive	.117
	Negative	-.350
Kolmogorov-Smirnov Z		2.711
Asymp. Sig. (2-tailed)		.000
Exact Sig. (2-tailed)		.000
Point Probability		.000

a. Test distribution is Uniform.

7.5 两个独立样本的检验

在第 1 章 1.2.6 的成组设计中, 取得的两个独立样本均服从正态分布时, 两组间均值

差异的比较使用 T 检验。但有时样本所隶属总体的分布类型可能不明或是非正态的, 此时, 两个样本之间是否有差异的检验就转化为两个独立样本间是否具有相同分布的检验。需要用到本节中的两个独立样本检验 (Two Independent Samples Test) 的方法。

7.5.1 曼-惠特尼 U 检验和威尔科克森秩和检验

曼-惠特尼 U (Mann-Whitney U) 检验和威尔科克森 (Wilcoxon) 秩和检验这两个方法都是用来分析两个独立样本同分布的检验, 它们是等价的。

1. 秩的计算

设样本 x_1, x_2, \dots, x_m 和样本 y_1, y_2, \dots, y_n 分别抽自相互独立的连续型随机变量总体 $F(x)$ 和 $G(x)$ 。并设这两个样本的合并样本 $x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_n$ 的各个单元之间互不相等, 则合并的样本容量 $N = m + n$ 。

对合并后的样本, 按从小到大的顺序排列, 并以 R_i 记 y_i 在混合样本的顺序统计量中的位置, 称 R_i 为 y_i 在混合样本的秩。则 Y 样本 y_1, y_2, \dots, y_n 的秩和 $W_y = \sum_{i=1}^n R_i$, 同样地,

以 Q_i 记 x_i 在混合样本的顺序统计量中的位置, 则 X 样本 x_1, x_2, \dots, x_m 的秩和 $W_x = \sum_{i=1}^m Q_i$ 。

由此可知, $W_x + W_y = 1 + 2 + \dots + N = N(N+1)/2$ 。因此, 用 W_x 作为统计量和用 W_y 作为统计量是相互等价的。

2. 假设检验

所要作的原假设为 $H_0: F(x) = G(x)$, 在原假设为真时, 若 $\min\{m, n\} \rightarrow \infty$, 且 $m/N \rightarrow \lambda \in (0, 1)$, λ 是一个常数, 则威尔科克森 (Wilcoxon) 秩和统计量 W_y 的概率分布和累积概率分布分别为

$$P(W_y = d) = \frac{t_{m,n}(d)}{\binom{N}{n}}$$

$$P(W_y \geq d) = \frac{\sum_{i \leq d} t_{m,n}(i)}{\binom{N}{n}}, d = n(n+1)/2, \dots, n(n+1)/2 + mn$$

其中, $t_{m,n}(d)$ 表示从 $1, 2, \dots, N = m + n$ 这 N 个数中任取 n 个数, 其和恰为 d 的取法种数。

W_y 的渐近正态性简记为

$$W_y \sim N(n(N+1)/2, mn(N+1)/12)$$

故

$$U = \frac{W_y - n(N+1)/2}{\sqrt{mn(N+1)/12}} \xrightarrow{L} N(0,1)$$

当观察值中有相等的值, 即有结时, 需通过对相等的观察值取平均秩, 来对威尔科克森 (Wilcoxon) 秩和统计量 W_y 修正, 此时, W_y 的渐近正态性为

$$W_y \sim N\left(n(N+1)/2, mn(N+1)/12 - nm \sum_{i=1}^g (t_i^3 - t_i)/(12N(N-1))\right)$$

其中, t_i 为结的长度, $i=1, 2, \dots, g$ 。

曼-惠特尼 (Mann-Whitney U) 统计量为

$$W_{xy} = W_y - n(n+1)/2, \quad W_{yx} = W_x - m(m+1)/2$$

它与威尔科克森 (Wilcoxon) 秩和统计量 W_y 只相差一个常数 $n(n+1)/2$ 。

同样在满足上述条件下, W_{xy} 的概率分布和累积概率分别为

$$P(W_{xy} = d) = P(W_y = d + n(n+1)/2) = \frac{t_{m,n}(d + n(n+1)/2)}{\binom{N}{n}}$$

$$P(W_{xy} \leq d) = P(W_y \leq d + n(n+1)/2) = \frac{\sum_{i \leq d} t_{m,n}(i + n(n+1)/2)}{\binom{N}{n}}, d = 0, 1, \dots, mn$$

W_{xy} 的渐近正态性简记为

$$W_{xy} \sim N(mn/2, mn(N+1)/12)$$

$$U = \frac{W_{xy} - mn/2}{\sqrt{mn(N+1)/12}} \xrightarrow{L} N(0,1), m, n \rightarrow \infty$$

同样,在有结取平均秩时,需对其方差作修正,此时, W_{xy} 的渐近正态性简记为

$$W_{xy} \sim N\left(mn/2, mn(N+1)/12 - nm \sum_{i=1}^g (t_i^3 - t_i)/(12N(N-1))\right)$$

当 $P(H_0) < \alpha$ 时,拒绝原假设。

3. 实例

例 7.10 为比较两种型号的汽车每加仑汽油的行驶里程,随机从每个型号的汽车中各抽取 12 辆汽车,并让每辆汽车高速行驶 1000 英里,得到每辆汽车每加仑汽油行驶里程数据,见表 7-17。试分析两种型号汽车每加仑汽油的行驶里程是否是相同?

表 7-17 两种型号的汽车每加仑汽油的行驶里程记录表

型号 1	20.6	19.9	18.6	18.9	18.8	20.2	24.2	22.5	19.8	19.8	19.2	20.5
型号 2	21.3	17.6	17.4	18.5	19.7	21.1	17.3	18.8	17.8	14.2	18.0	20.1

在 SPSS 中,解题步骤如下:

(1) 在 SPSS 数据编辑窗口中,建立数据文件。
具体内容参见 data07-13.sav。

(2) 按 Analyze→Nonparametric Tests→2 Independent Samples 顺序单击菜单项,展开 Two Independent Samples 对话框,如图 7-6 所示。

(3) 指定检验变量。从变量列表中选择 **行驶里程** 变量,移到 Test Variable List 框中。

指定分组变量名。从左面变量列表中选择 **型号** 变量,并移到 Grouping Variable 框中,单击 Define Groups 按钮,展开如图 7-7 所示的对话框,在两个编辑栏中输入分组值 1 和 2。

确定用来进行检验的方法。在 Test Type 框中, Mann-Whitney U。

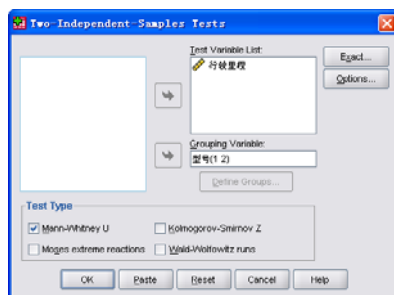


图 7-6 两个独立样本检验对话框

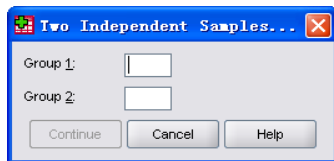


图 7-7 定义分组对话框

(4) 单击 Exact 按钮,在弹出的选项卡中,选择 Exact 选项,单击 Continue 按钮返回 Two Independent Samples 对话框。

(5) 单击 OK 按钮,在输出窗中得到计算结果,见表 7-18 和表 7-19。

(6) 结果和讨论

从表 7-18 可见,第 1 种型号汽车的每加仑汽油的行驶里程的平均秩为 15.79,秩和为

189.50, 第 2 种型号汽车的每加仑汽油的行驶里程的平均秩为 9.21, 秩和为 110.50。

表 7-18 秩和表

Ranks				
	型号	N	Mean Rank	Sum of Ranks
行驶里程	第1种型号	12	15.79	189.50
	第2种型号	12	9.21	110.50
	Total	24		

表 7-19 秩检验表

Test Statistics ^a	
	行驶里程
Mann-Whitney U	32.500
Wilcoxon W	110.500
Z	-2.282
Asymp. Sig. (2-tailed)	.023
Exact Sig. [2*(1-tailed Sig.)]	.020 ^a
Exact Sig. (2-tailed)	.021
Exact Sig. (1-tailed)	.011
Point Probability	.001

a. Not corrected for ties.

b. Grouping Variable: 型号

从表 7-19 的秩检验表可见, Mann-Whitney U 统计量为 32.500, Wilcoxon 秩和统计量 W_y 为 110.50, 计算得到的 Z 值为 -2.282, 渐近分布双侧检验的概率为 0.023, 在未校正结时, 单侧精确检验的概率为 0.020, 在校正结时, 双侧精确检验的概率为 0.021, 单侧精确检验的概率为 0.011, 在 $W_{xy} = 32.500$ 点处的概率为 0.001。

由于在两种型号汽车每加仑汽油的行驶里程相同原假设下, 出现目前统计量的值或者更极端值的双侧检验概率为 0.021 小于 0.05, 故拒绝原假设, 认为两种汽车型号的每加仑汽油的行驶里程是不相同的, 据于第 1 种型号的平均秩高于第 2 种型号, 可以认为第 1 种型号汽车在高速状态下要比第 2 种型号汽车省油。

7.5.2 柯尔莫哥洛夫-斯米诺夫检验 Z 检验

1. 柯尔莫哥洛夫-斯米诺夫检验 Z (Kolmogorov-Smirnov Z) 检验的基本思想

Kolmogorov-Smirnov Z , 是更普通的探测两者在位置和分布形状上差异的检验。该检验是建立在两个样本的累积分布函数之间的最大绝对差异的基础上的。当这个差异显著地大时, 两个分布被认为是存在差异的。

设样本 x_1, x_2, \dots, x_{n_1} 和样本 y_1, y_2, \dots, y_{n_2} 分别抽自相互独立的连续型随机变量总体 $F_1(x)$ 和 $F_2(x)$, $\hat{F}_1(x)$ 和 $\hat{F}_2(x)$ 分别是两个样本的对应累积经验分布函数。

经验分布函数和差异的计算:

两组数据分别地按有小到大的顺序排列成, $X_{[1]}$ 到 $X_{[n]}$, 第 i 组的经验累积函数用下式计算

$$\hat{F}_i(X) = \begin{cases} 0 & -\infty < X < X_{[1]} \\ j/n_i & X_{[j]} \leq X < X_{[j+1]} \\ 1 & X_{[n]} \leq X < \infty \end{cases}$$

对两组中所有的 X_j 值, 两组间的差异为

$$D_j = \hat{F}_1(X_j) - \hat{F}_2(X_j)$$

其中, $\hat{F}_1(X_j)$ 是对应于较大样本含量组的累积概率函数。同时计算最大正值、负值和绝对值。

所要检验的原假设为

$$H_0: F_1(x) = F_2(x)$$

Kolmogorov-Smirnov 提出的统计量为

$$D_{n_1, n_2} = \sup_{-\infty < x < \infty} |\hat{F}_1(x) - \hat{F}_2(x)|$$

当 H_0 为真时, $Z = \sqrt{n}D_{n_1, n_2}$ 有极限分布, 其中 $n = n_1 n_2 / (n_1 + n_2)$ 。其渐近分布参见 7.4.1。

当 $P(H_0) < \alpha$ 时, 拒绝原假设。

2. 实例

例 7.11 从用两种材料的灯丝制造的灯泡中, 分别随机抽取一个样本, 得到如下灯泡使用寿命的样本数据, 见表 7-20, 试用 Kolmogorov-Smirnov Z 法检验两种材料制成的灯泡的使用寿命有无显著性差异?

表 7-20 两种材料制成的灯泡的使用寿命 (单位: 小时)

甲	1610	1650	1680	1700	1750	1720	1800
乙	1580	1600	1640	1650	1700		

(1) 在 SPSS 数据编辑窗口中, 建立数据文件。具体内容参见 data07-14.sav。

(2) 按 Analyze→Nonparametric Tests→2 Independent Samples 顺序单击菜单项, 展开 Two Independent Samples 对话框, 如图 7-6 所示。

(3) 指定检验变量。从变量列表中选择 *寿命* 变量, 移到 Test Variable List 框中。

指定分组变量名。从左面变量列表中选择 *材料* 变量, 并移到 Grouping Variable 框中, 单击 Define Groups 按钮, 展开如图 7-7 所示的对话框, 在两个编辑栏中输入分组值 1 和 2。

确定用来进行检验的方法。在 Test Type 框中, Kolmogorov-Smirnov Z。

(4) 单击 Exact 按钮, 在弹出的选项卡中, 选择 Monte Carlo 选项, 单击 Continue 按钮, 返回 Two Independent Samples 对话框。

(5) 单击 OK 按钮, 在输出窗中得到计算结果, 见表 7-21 和表 7-22。

(6) 结果和讨论

表 7-21 给出了参与检验的两组的样本量, 由表 7-21 可见, 甲组的样本量为 7, 乙组的样本量为 5。

表 7-21 样本量表

Frequencies		
寿命	材料	N
	甲	7
	乙	5
	Total	12

表 7-22 检验统计表

Test Statistics ^a			寿命
Most Extreme Differences	Absolute		.514
	Positive		.000
	Negative		-.514
Kolmogorov-Smirnov Z			.878
Asymp. Sig. (2-tailed)			.423
Monte Carlo Sig. (2-tailed)	Sig.		.305 ^a
	99% Confidence Interval	Lower Bound	.293
		Upper Bound	.317

a. Based on 10000 sampled tables with starting seed 2000000.

b. Grouping Variable: 材料

表 7-22 列出了检验过程中用到的统计量。从上到下依次为最大偏差的绝对值 (0.514)、正值 (0)、负值 (-0.514), Kolmogorov-Smirnov Z 统计量 (0.878), 双侧检验的渐近分布的概率为 0.423, 双侧检验的 Monte Carlo 法的概率为 0.305, Monte Carlo 法的概率的 99% 的置信区间的下限为 0.293, 上限为 0.317。

由于本例的样本量较小, 故采用 Monte Carlo 法的概率来判定, 在两种材料制成的灯泡的使用寿命相同的原假设下, 出现目前统计量的值或者更极端值的双侧检验概率为 0.305 大于 0.05, 故现有证据不足以支持拒绝原假设。认为两种材料制成的灯泡的使用寿命无显著性差异。

7.5.3 摩西极端值反应检验

摩西极端值反应 (Moses Extreme Reactions) 检验, 假设实验变量影响某个方向上的一些被试对象和相反方向上的其他被试对象。它检验同控制组比较的极端响应。本检验关键是控制组的跨度, 以及当合并控制组时, 实验组里有多少极端值影响跨度。

1. 跨度计算

设样本 x_1, x_2, \dots, x_{n_1} 和样本 y_1, y_2, \dots, y_{n_2} 分别抽自相互独立的连续型随机变量总体 $F_1(x)$ 和 $F_2(x)$ 。来自两组的观察值一起排序和赋秩, 相同观察值被平均赋秩。确定对应于最小和最大控制组 (第一组) 的秩成员, 跨度用下式计算:

跨度 (SPAN) = 秩 (控制组中的最大值) - 秩 (控制组中的最小值) + 1,
取整到最接近的整数。

2. 显著性水平

所要检验的原假设为:

$$H_0: F_1(x) = F_2(x)$$

精确的单侧概率水平用下式计算

$$P(SPAN \leq n_c - 2h + g) = \sum_{i=0}^g \left[\binom{i + n_c - 2h - 2}{i} \binom{n_e + 2h + 1 - i}{n_e - i} \right] / \binom{n_c + n_e}{n_c}$$

其中, $h=0$, n_c 是控制组中样品的数量, n_e 是实验组中样品的数量。同样的公式在下一步中使用, 那儿 h 不等于 0。

3. 极差的审查

重复上面的检验, 从控制组中去掉最小的 h 和最大的 h 秩。由于跨度范围很容易受偶然因素的影响, 所以两端 5% 的样品被自动地删除。如果用户不指定, h 采用 $0.05n_c$ 的整数部分或 1 中的较大者, 如果用户指定, 除非它小于 1, 否则使用用户指定的整数值。显著性水平由上面 2 中确定。

4. 实例

例 7.12 为了鉴别甲、乙两厂生产的同种肥料的质量, 分别从两厂生产的肥料中各随机抽取一个样本, 测得有效成分的含量分别为:

甲: 7, 52, 49, 14, 22, 36, 40, 48, 36, 27, 19

乙: 39, 12, 21, 24, 9, 4, 17, 7, 10, 18, 20, 5

能否判断两厂的肥料有显著性差异?

(1) 在 SPSS 数据编辑窗口中, 建立数据文件。具体内容参见 data07-15.sav。

(2) 按 Analyze→Nonparametric Tests→2 Independent Samples 顺序单击菜单项, 展开 Two Independent Samples 对话框, 如图 7-6 所示。

(3) 指定检验变量。从变量列表中选择含量变量, 移到 Test Variable List 框中。

指定分组变量名。从左面变量列表中选择厂名变量, 并移到 Grouping Variable 框中, 单击 Define Groups 按钮, 展开如图 7-7 所示的对话框, 在两个编辑栏中输入分组值 1 和 2。

确定用来进行检验的方法。在 Test Type 框中, 选择 Moses Extreme Reactions。

(4) 单击 Exact 按钮, 在弹出的选项卡中, 选择 Monte Carlo 选项, 单击 Continue 按钮, 返回 Two Independent Samples 对话框。

(5) 单击 OK 按钮, 在输出窗中得到计算结果, 见表 7-23 和表 7-24。

(6) 结果和讨论

表 7-23 给出了参与检验的两组的样本量, 由表 7-23 可见, 甲组(控制组)的样本量为 11, 乙组(实验组)的样本量为 12。

表 7-24 列出了检验中的各个统计量, 从上到下依次为观察到的控制组的跨度为 21,

单侧检验原假设成立的概率为 0.534，截了两侧部分的极端值后的跨度为 15，单侧检验的概率为 0.247。从每侧各截掉的异常值数为 1。

表 7-23 样本量表

Frequencies		
含量	厂名	N
	甲 (Control)	11
	乙 (Experimental)	12
	Total	23

表 7-24 检验统计表

Test Statistics ^{a, b}		含量
Observed Control Group		21
Span	Sig. (1-tailed)	.534
Trimmed Control Group		15
Span	Sig. (1-tailed)	.247
Outliers Trimmed from each End		1

a. Moses Test

b. Grouping Variable: 厂名

由于 P 为 0.247，大于 0.05，故不拒绝原假设，而认为两厂的肥料没有显著性差异。

7.5.4 沃尔德-乌尔夫威兹游程检验

沃尔德-乌尔夫威兹游程 (Wald-Wolfowitz Runs) 检验是更普通的探测两者在位置和分布形状上差异的检验。Wald-Wolfowitz 游程检验对两组观察合并和赋予秩。如果两个样本是来自相同的总体，则合并后的两组应有较多的游程数。

1. 计算游程的数量

设样本 x_1, x_2, \dots, x_{n_1} 和样本 y_1, y_2, \dots, y_{n_2} 分别抽自相互独立的连续型随机变量总体 $F_1(x)$ 和 $F_2(x)$ ，合并两个样本的所有观察值并由小到大排列为升序。计算对应于有序数据中同组数字变化的次数。游程数 (R) 等于同组数字变化的次数加 1。

如果两组观察数据中包含结，计算可能的游程最小数量和最大数量。

2. 显著性水平

所作的原假设为： $H_0: F_1(x) = F_2(x)$ 。

如果 $n_1 + n_2$ ，总的样本量，小于或等于 30，单侧显著性水平可由下式中精确地计算：

$$\text{当 } R \text{ 是偶数时, } P(r \leq R) = \frac{2}{\binom{n_1 + n_2}{n_1}} \sum_{r=2}^R \binom{n_1 - 1}{r/2 - 1} \binom{n_2 - 1}{r/2 - 1}$$

$$\text{当 } R \text{ 是奇数时, } P(r \leq R) = \frac{1}{\binom{n_1 + n_2}{n_1}} \sum_{r=2}^R \binom{n_1 - 1}{k - 1} \binom{n_2 - 1}{k - 2} \binom{n_1 - 1}{k - 2} \binom{n_2 - 1}{k - 1}$$

其中, $r = 2k - 1$ 。

对于样本量大于 30, 使用正态渐近分布。参见游程检验。

3. 实例

例 7.13 有两组少儿肝炎患者, 一组为甲、乙型肝炎病毒混合感染, 另一组为重叠感染, 治疗者的治愈天数的数据存放在 data07-16.sav 的数据文件中, 试用 Wald-Wolfowitz Runs 秩和检验法检验两组治疗之间有无显著性差异?

(1) 在 SPSS 数据编辑窗口中, 打开 data07-16.sav。

(2) 按 Analyze→Nonparametric Tests→2 Independent Samples 顺序单击菜单项, 展开 Two Independent Samples 对话框, 如图 7-6 所示。

(3) 指定检验变量。从变量列表中选择治愈天数变量, 移到 Test Variable List 框中。

指定分组变量名。从左面变量列表中选择组别变量, 并移到 Grouping Variable 框中, 单击 Define Groups 按钮, 展开如图 7-7 所示的对话框, 在两个编辑栏中输入分组值 1 和 2。

确定用来进行检验的方法。在 Test Type 框中, Wald-Wolfowitz Runs。

(4) 单击 Ok 按钮, 在输出窗中得到计算结果, 见表 7-25 和表 7-26。

(5) 结果和讨论

表 7-25 给出了参与检验的两组的样本量, 由表 7-25 可见, 混合感染组的样本量为 41, 重叠感染组的样本量为 24。

表 7-25 样本量表

Frequencies		
组别		N
治愈天数	混合感染	41
	重叠感染	24
Total		65

表 7-26 检验统计表

Test Statistics ^{a,c}				
		Number of Runs	Z	Asymp. Sig. (1-tailed)
治愈天数	Minimum Possible	24 ^a	-1.955	.025
	Maximum Possible	32 ^a	.194	.577

a. There are 5 inter-group ties involving 14 cases.

b. Wald-Wolfowitz Test

c. Grouping Variable: 组别

表 7-26 列出了检验统计量, 由于在混合样本观察值中有 14 个样品出现了 5 个结, 所以应选用最小可能游程的检验结果, 最小可能游程数为 24 个, Z 值为-1.955 (合并样本总含量大于 30), 单侧检验的概率 0.025, 小于 0.05, 故拒绝原假设。认为两组治疗之间有显著性差异。

7.6 多个独立样本的检验

在第 1 章 1.2.6 的单因素多水平设计中, 取得的多个独立样本均服从正态分布时, 多组间均值差异的比较使用单因素方差分析。但有时样本所隶属总体的分布类型可能不明

或是非正态的, 此时, 多个样本之间是否有差异的检验就转化为多个独立样本间是否具有相同分布的检验。需要用到本节中的多个独立样本检验 (K Independent Samples Test) 的方法。

7.6.1 克鲁斯卡-沃里斯 H 检验

克鲁斯卡-沃里斯 H (Kruskal-Wallis H) 检验是曼-惠特尼 U (Mann-Whitney U) 检验的扩展, 类似于单方向方差分析, 探究分布位置上的差异。该方法假设从 k 个无序的总体中抽取样本。它所检验的问题称为无方向检验问题。

1. 秩和的计算

合并 K 个非空组中的观察值, 排序并赋秩, 有结时对其取平均秩, 并求得结的长度 t_i , 以及计算 $T_i = t_i^3 - t_i$ 的总和。获取每一组的秩和 R_i 和观察值的数量 n_i 。

2. 检验统计量和显著性水平

对于有结时, 不校正的检验统计量为

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k R_i^2 / n_i - 3(N+1)$$

其中, $N = \sum_{i=1}^k n_i$

对于有结时, 校正的统计量为

$$H' = \frac{H}{1 - \sum_{i=1}^m T_i / (N^3 - N)}$$

其中, m 为结集的数量。

所要检验的原假设为 $H_0: \theta_1 = \theta_2 = \dots = \theta_k$, 即 k 个位置参数相等。当原假设为真时, 统计量 H 和 H' 渐近服从 $\chi^2(k-1)$ 分布。故显著性水平基于自由度为 $K-1$ 的 χ^2 分布。

当 $P(H_0) < \alpha$ 时, 拒绝原假设。

3. 实例

例 7.14 某医院妇产科为考察几种卵巢功能异常患者血清中促黄体素的水平, 测试了 3 组患者血清中促黄体素的水平, 测试结果数据存放在 data07-17.sav 的数据文件中, 试用 Kruskal-Wallis H 法检验这 3 组患者血清中促黄体素的水平有无差异?

在 SPSS 中的解题步骤如下:

- (1) 在 SPSS 数据编辑窗口中，打开 data07-17.sav。
- (2) 按 Analyze→Nonparametric Tests→K Independent Samples 顺序展开 K Independent Samples 对话框。见图 7-8。

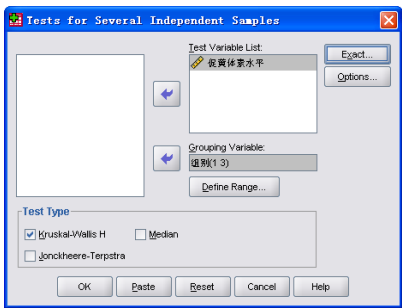


图 7-8 K Independent Samples 对话框

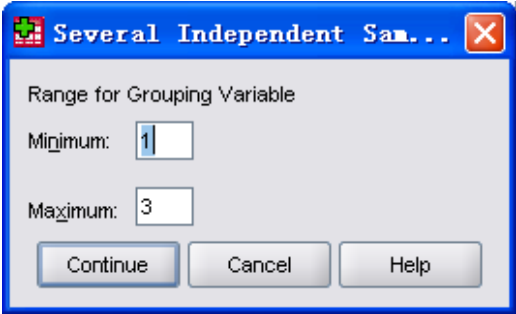


图 7-9 Define Range 对话框

- 选择促黄体素水平变量进入 Test Variable List 框。
- 选择组别变量进入 Grouping Variable 框中。单击 Define Range 按钮，展开 Define Range 对话框，见图 7-9，在 Minimum 框中输入 1，在 Maximum 框中输入 3。
- 单击 Continue 按钮返回 K Independent Samples 对话框。
- (3) 在 Test Type 框选择 Kruskal-Wallis 方法。
 - (4) 单击 OK 按钮，提交运算。在输出窗口中，得到输出结果，见表 7-27、表 7-28。
 - (5) 结果与讨论

表 7-27 平均秩表

Ranks			
组别		N	Mean Rank
促黄体素水平	卵巢发育不良	5	22.00
	丘脑性闭经	11	9.00
	垂体性闭经	8	11.38
	Total	24	

表 7-28 检验统计表

Test Statistics ^{a,b}	
	促黄体素水平
Chi-Square	11.922
df	2
Asymp. Sig.	.003

a. Kruskal Wallis Test
b. Grouping Variable: 组别

在表 7-27 中，列出了分组的名称、各组的样本量和平均秩，其中平均秩最高的组为卵巢发育不良组为 22.00，丘脑性闭经组的平均秩最高平均秩最低为 9。

在表 7-28 中可见，Kruskal-Wallis H 检验中的 χ^2 值为 11.922，自由度为所分组数减 1 等于 2，在 3 组患者血清中促黄体素的水平无差异的原假设下，出现目前统计量的值或者更极端值的双侧检验概率为 0.003，小于 0.05，故可认为这 3 组患者血清中促黄体素的水平有显著性差异。结合表 7-27 可知，三组中卵巢发育不良组的血清中促黄体素的水平最

高、丘脑性闭经组的血清中促黄体素的水平最低，而垂体性闭经组的血清中促黄体素的水平居中。

7.6.2 中位数检验

中位数 (Median) 检验，是很普通的检验，但效率不高，探究在位置和形状上的分布差异。该方法假设从 k 个无序的总体中抽取样本。它也是用来检验无方向问题的。

1. 表格构建

如果中位数未被使用者指定，则它通过所有组合并数据排序后按下面公式计算确定

$$M_d = \begin{cases} (X_{[N/2]} + X_{[N/2+1]})/2 & \text{如果 } N \text{ 是偶数} \\ X_{[(N+1)/2]} & \text{如果 } N \text{ 是奇数} \end{cases}$$

其中， $X_{[M]}$ 为最大值， $X_{[1]}$ 是最小值。

各组统计超过中位数和小于等于中位数的样品数量，则得到以下的表格，见表 7-29。

表 7-29 用中位数得到的各组分类表

组别						
	1	2	3	...	k	
小于等于中位数	O_{11}	O_{12}	O_{13}	...	O_{1k}	R_1
大于中位数	O_{21}	O_{22}	O_{23}	...	O_{2k}	R_2
	n_1	n_2	n_3	...	n_k	N

表中的 R_i 表示行和， n_i 为列和。

2. 检验

所有非空组的 χ^2 统计量用下式计算

$$\chi^2 = \sum_{j=1}^k \sum_{i=1}^2 (O_{ij} - E_{ij})^2 / E_{ij}$$

其中， $E_{ij} = \frac{R_i n_j}{N}$

所要检验的原假设为： $H_0: M_{d_1} = M_{d_2} = \cdots = M_{d_k}$ ，在 H_0 为真时，上式确定的 χ^2 近

似服从 $\chi^2(k-1)$ 的卡方分布。

当 $P(H_0) < \alpha$ 时，拒绝原假设。

3. 实例

例 7.15 将三种不同菌型的伤寒病毒 A_1, A_2, A_3 分别接种于三组同类型的小白鼠上，观察它们的存活天数，测得的数据存放在 data07-18.sav 的数据文件中，试用 Median（中位数）法检验三种不同菌型下小白鼠存活天数上有无显著性差异？

在 SPSS 中的解题步骤如下：

(1) 在 SPSS 数据编辑窗口中，打开 data07-18.sav。

(2) 按 Analyze→Nonparametric Tests→K Independent Samples 顺序展开 K Independent Samples 对话框。见图 7-8。

选择存活天数变量进入 Test Variable List 框。

选择细菌类型变量进入 Grouping Variable 框中。单击 Define Range 按钮，展开 Define Range 对话框，见图 7-9，在 Minimum 框中输入 1，在 Maximum 框中输入 3。

单击 Continue 按钮返回 K Independent Samples 对话框。

在 Test Type 框选择 Median 方法。

(3) 单击 Exact 按钮，在弹出的对话框中，选择 Exact 选项，单击 Continue 按钮返回 K Independent Samples 对话框。

(4) 单击 OK 按钮，提交运算。在输出窗口中，得到输出结果，见表 7-30、表 7-31。

(5) 结果与讨论

表 7-30 中位数分类表

Frequencies			
	细菌类型		
	A1	A2	A3
存活天数 > Median	2	4	6
≤ Median	8	5	5

表 7-31 检验统计量表

Test Statistics ^a	
	存活天数
N	30
Median	6.0000
Chi-Square	2.710 ^a
df	2
Asymp. Sig.	.258
Exact Sig.	.298
Point Probability	.061

a. 3 cells (50.0%) have expected frequencies less than 5. The minimum expected cell frequency is 3.6.
b. Grouping Variable: 细菌类型

表 7-30 列出了各组大于中位数和小于等于中位数的观察值的个数，从该表可见，由于各组的观察次数均不多，故本例需要用精确检验。

表 7-31 列出了检验中的各种统计量，从上到下依次为总的观察次数为 30，中位数为 6， $\chi^2 = 2.710$ ，自由度为 2，在三种不同菌型下小白鼠存活天数上无差异的原假设下，出现目前统计量的值或者更极端值的渐近分布双侧检验概率为 0.258，精确检验的概率为

0.298, 点概率为 0.061。

备注 a 说明有 50% 的期望频数小于 5, 最小频数为 3.6, 故本例要用精确检验的结果, 由于精确检验下的概率 0.298 大于 0.05, 故现有证据不足以拒绝原假设, 即认为三种不同菌型下小白鼠存活天数上无显著性差异。

7.6.3 乔卡契尔-特普斯特拉检验

乔卡契尔-特普斯特拉 (Jonckheere-Terpstra) 检验。当 k 个总体有序 (升序或降序) 时, 即检验的问题有方向性时, 此检验方法是非常有效的。例如, k 个总体可以描述 k 个增加的温度。检验的假设是不同的温度产生同样响应的分布, 备择假设: 温度升高响应剧烈。这里, 假设两个样本是有序的, 因此, 使用 Jonckheere-Terpstra 检验是最适当的。在 SPSS 中, 只有安装了 Exact Tests 时, 此检验才是可选用的。

1. 乔卡契尔-特普斯特拉检验的基本思想

设有 k 个连续型随机变量总体: X_1, X_2, \dots, X_k , $x_{i1}, x_{i2}, \dots, x_{in_i}$ 是来自第 i 个总体 X_i 的样本, 其样本量为 n_i , $i = 1, 2, \dots, k$, 则总的样本容量为 $N = \sum_{i=1}^k n_i$ 。所有 N 个样本单元都是相互独立的。

设第 i 个总体 X_i 的分布函数为 $F(x - \theta_i)$, $i = 1, 2, \dots, k$ 。对于有方向性的检验问题, 所要检验的原假设 $H_0: \theta_1 = \theta_2 = \dots = \theta_k$, $H_1: \theta_1 \leq \theta_2 \leq \dots \leq \theta_k$, 且 $\theta_1 < \theta_k$ 。

Jonckheere-Terpstra 检验用 J 作为统计量, $J = \sum_{1 \leq i < j \leq k} W_{ij}$, 其中的 W_{ij} 就是 Mann-Whitney U 统计量。 $W_{ij} = \# \{ (x_{ir}, x_{js}) : x_{ir} < x_{js}, r = 1, 2, \dots, n_i; s = 1, 2, \dots, n_j \}$ 。式中的“#”表示计数。

可以证明, 在原假设为真时, 若 $\min \{n_1, \dots, n_k\} \rightarrow \infty$, 且对所有的 $i = 1, 2, \dots, k$, 都有 $n_i / N \rightarrow \lambda_i \in (0, 1)$, 则

$$J - T = \frac{J - E(J)}{\sqrt{D(J)}} = \frac{J - \frac{1}{4} \left[N^2 - \sum_{i=1}^k n_i^2 \right]}{\sqrt{\frac{1}{72} \left[N^2 (2N + 3) - \sum_{i=1}^k n_i^2 (2n_i + 3) \right]}} \xrightarrow{L} N(0, 1)$$

当全部样本中结的个数为 g 时, 需对上式中的 $D(J)$ 部分做如下修正

$$\begin{aligned}
D(J) = & \frac{1}{72} \left[N^2(2N+3) - \sum_{i=1}^k n_i^2(2n_i+3) - \sum_{s=1}^g t_s(t_s-1)(2t_s+5) \right] \\
& + \frac{1}{36N(N-1)(N-2)} \left[\sum_{i=1}^k n_i(n_i-1)(n_i-2) \right] \left[\sum_{s=1}^g t_s(t_s-1)(t_s-2) \right] \\
& + \frac{1}{8N(N-1)} \left[\sum_{i=1}^k n_i(n_i-1) \right] \left[\sum_{s=1}^g t_s(t_s-1) \right]
\end{aligned}$$

当 $P(H_0) < \alpha$ 时, 拒绝原假设。

2. 实例

例 7.16 随机选择 45 个患者, 将他们随机地分成三组, 每组 15 人, 分别服用 2.4 克、4.8 克和 7.2 克的降血脂新药。6 周后测得每个被试对象的低密度脂蛋白胆固醇含量, 数据存放在 data07-19.sav 的数据文件中, 试用 Jonckheere-Terpstra 法检验有没有这样的趋势, 即服用 7.2 克的降血脂新药的低密度脂蛋白胆固醇含量最低, 服用 4.8 克的降血脂新药的低密度脂蛋白胆固醇含量居中, 而服用 2.4 克的降血脂新药的低密度脂蛋白胆固醇含量最高?

本例是一个检验有方向性的问题, 适合用 Jonckheere-Terpstra 法检验。

在 SPSS 中的解题步骤如下:

(1) 在 SPSS 数据编辑窗口中, 打开 data07-19.sav。

(2) 按 Analyze→Nonparametric Tests→K Independent Samples 顺序展开 K Independent Samples 对话框。见图 7-8。

选择低密度脂蛋白胆固醇含量变量进入 Test Variable List 框。

选择用药量类别变量进入 Grouping Variable 框中。单击 Define Range 按钮, 展开 Define Range 对话框, 见图 7-9, 在 Minimum 框中输入 1, 在 Maximum 框中输入 3。

单击 Continue 按钮返回 K Independent Samples 对话框。

在 Test Type 框选择 Jonckheere-Terpstra 方法。

(3) 单击 OK 按钮, 提交运算。在输出窗口中, 得到输出结果, 见表 7-32。

(4) 结果与讨论

表 7-32 显示了 Jonckheere-Terpstra 检验结果, 从上到下各行列出的依次为用药量类

表 7-32 Jonckheere-Terpstra 表

Jonckheere-Terpstra Test ^a	
Number of Levels in 用药量类别	3
N	45
Observed J-T Statistic	493.500
Mean J-T Statistic	337.500
Std. Deviation of J-T Statistic	48.011
Std. J-T Statistic	3.249
Asymp. Sig. (2-tailed)	.001

a. Grouping Variable: 用药量类别

别/变量中的水平数为 3, 总样本量为 45, J 统计量值为 493.5, $E(J) = 337.5$, $D(J) = 48.011^2$, $J - T = 3.249$, 双侧检验的概率为 0.001。

由于在服用不同降血脂新药量会产生相同低密度脂蛋白胆固醇含量的原假设下, 出现目前统计量的值或者更极端值的双侧检验概率为 0.001, 小于 0.01, 故拒绝原假设, 而认为服用不同降血脂新药量会产生不同低密度脂蛋白胆固醇含量, 即服用 7.2 克的降血脂新药后, 低密度脂蛋白胆固醇含量最低, 服用 4.8 克的降血脂新药后, 低密度脂蛋白胆固醇含量居中, 而服用 2.4 克的降血脂新药后的低密度脂蛋白胆固醇含量最高。

7.7 两个相关样本检验

在第 1 章 1.2.6 中, 对于从配对试验中取得的容量相等的两组样本数据, X_1, X_2, \dots, X_N 和 Y_1, Y_2, \dots, Y_N , 由于同一个样本数据之间, 可以假设为相互独立, 但不能假设它们同分布, 所以不能用上面两个独立样本的非参数检验方法来处理。应根据数据资料的连续型和离散型的分类, 分别用以下介绍的四种方法中最适宜的一种来进行处理。

7.7.1 威尔科克森检验

威尔科克森 (Wilcoxon) 符号等级检验, 适用于连续型数据, 它不但考虑两个符号的差异, 而且还考虑成对样品数值之间差异。所以, 它比下一节介绍的符号检验更有效。

威尔科克森 (Wilcoxon) 符号等级检验的基本做法如下:

1. 秩差异的计算

对于每个样品, 在排序前, 先计算成对观察值之间的差异 $D_i = X_i - Y_i$ 和 $|D_i|$ 。将所有非零的绝对差排列成升序并赋秩。在有结的情况下, 对结点使用平均秩。计算对应于正差异的秩和 S_p 和负差异的秩和 S_n 。正秩的均值为

$$\bar{X}_p = S_p / n_p$$

负秩的均值为

$$\bar{X}_n = S_n / n_n$$

其中, n_p 是有正差异的样品的数量, n_n 是有负差异的样品的数量。

2. 检验

$$H_0: \text{对称中心 } \theta = 0$$

在原假设为真时, 大样本下, 统计量

$$Z = \frac{\min(S_p, S_n) - (n(n+1)/4)}{\sqrt{n(n+1)(2n+1)/24 - \sum_{j=1}^l (t_j^3 - t_j)/48}} \xrightarrow{L} N(0,1)$$

其中, n 为非零差异样品的数量, l 结的数量, t_j 为第 j 个结的长度。

当 $P(H_0) < \alpha$ 时, 拒绝原假设。

3. 实例

例 7.17 甲、乙两人分析同一气体 CO_2 的含量, 得到的 CO_2 含量结果存放在 data07-20.sav 的数据文件中, 试用 Wilcoxon 符号等级检验法检验两人的分析结果间有无显著性差异?

在 SPSS 中的解题步骤如下:

(1) 在 SPSS 数据编辑窗口中, 打开 data07-20.sav。

(2) 按 Analyze→Nonparametric Tests→Two-Related-Samples Tests 顺序展开 Two-Related-Samples Tests 对话框。见图 7-10。

同时选中分析人员甲和分析人员乙变量进入 Test Pairs 下框中。

在 Test Type 框选择 Wilcoxon 方法。

(3) 根据样本量选择检验方法

由于本例为小样本, 故需要用精确检验。

单击 Exact 按钮, 在弹出的对话框中, 选择 Exact

选项, 单击 Continue 按钮返回 Two-Related-Samples Tests 对话框。

(4) 单击 OK 按钮, 提交运算。在输出窗口中, 得到输出结果, 见表 7-33、表 7-34。

(5) 结果与讨论

表 7-33 列出了分析人员甲测得的 CO_2 含量值小于分析人员乙测得的 CO_2 含量值的数量为 12 个, 平均秩为 11.04, 秩和为 132.50, 反之为 7 个, 平均秩为 8.21, 秩和为 57.50, 观察值相等数 (结) 为 1 个, 总观察值数为 20。

表 7-34 列出了 Wilcoxon 符号等级检验的统计量, 从上到下各行依次为: 根据正秩计算得到的 Z 值 -1.529, 双侧检验渐近分布计算得到的概率 0.126, 双侧检验用精确法计算得到的概率 0.136, 单侧检验用精确法计算得到的概率 0.068, 点概率为 0.008。

由于本例为小样本, 故用精确法计算结果, 在两人的分析结果间无差异的原假设下, 出现目前统计量的值或者更极端值的双侧检验概率为 0.136, 大于 0.05, 故没有足够的证



图 7-10 两个相关样本检验对话框

据支持拒绝原假设，暂时不能得出两人的分析结果间有显著性差异的结论。

表 7-33 Wilcoxon 符号等级检验表

Ranks				
		N	Mean Rank	Sum of Ranks
分析人员乙 - 分析人员甲	Negative Ranks	12 ^a	11.04	132.50
	Positive Ranks	7 ^b	8.21	57.50
	Ties	1 ^c		
	Total	20		

a. 分析人员乙 < 分析人员甲

b. 分析人员乙 > 分析人员甲

c. 分析人员乙 = 分析人员甲

表 7-34 Wilcoxon 符号等级检验统计量表

Test Statistics ^b	
	分析人员乙 - 分析人员甲
Z	-1.529 ^a
Asymp. Sig. (2-tailed)	.126
Exact Sig. (2-tailed)	.136
Exact Sig. (1-tailed)	.068
Point Probability	.008

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

7.7.2 符号检验

符号检验 (Sign test) 适用于连续型数据，符号检验对所有样品计算两个变量值间的差并将差值分为正、负或结 3 类。如果两个变量有类似的分布，则正、负数差异无显著不同。

符号检验的基本做法如下：

1. 符号合计

对于每个样品，计算差异 $D_i = X_i - Y_i$ ，合计正差异的数量 S_p 和负差异的数量 S_n ，对于 $X_i = Y_i$ 的样品，不算在正、负之列。

在满足一定的条件下， D_i 独立同分布，因此，两个样本间有无差异的检验问题，就等价于对称中心 θ 否等于 0 的检验问题。

2. 检验

所要做的原假设为 H_0 ：对称中心 $\theta = 0$ 。

如果 $n_p + n_n \leq 25$ ，当 $p = 0.5$ 和 $r = \min(n_p, n_n)$ 时，在 $n_p + n_n$ 次试验中， r 或少数“成功”事件的精确概率用下面的二项分布公式递推计算

$$p(X \leq r) = \sum_{i=0}^r \binom{n_p + n_n}{i} (0.5)^{(n_p + n_n)}$$

如果 $n_p + n_n > 25$ ，则显著性水平根据正态近似值

$$Z_c = \frac{\max(n_p, n_n) - 0.5(n_p + n_n) - 0.5}{0.5\sqrt{n_p + n_n}} \xrightarrow{L} N(0,1)$$

计算得到。

当 $P(H_0) < \alpha$ 时, 拒绝原假设。

3. 实例

例 7.18 某厂生产豪华型和普通型两种家用电器。从各个零售点抽样得到 10 个零售点的这两个电器的销售价格, 数据存放在 data07-21.sav 的数据文件中, 试用符号检验分析两种电器之间的差价是否满足厂家提出的相差 100 元的厂方建议?

在 SPSS 中的解题步骤如下:

1. 在 SPSS 数据编辑窗口中, 打开 data07-21.sav。

2. 利用第 2 章 2.5 中介绍的建立新变量的方法, 建立 **豪华型减 100** 新变量, 它等于 **豪华型-100**。

3. 按 Analyze→Nonparametric Tests→Two-Related-Samples Tests 顺序, 展开 Two-Related-Samples Tests 对话框。见图 7-10。

同时选中 **豪华型减 100** 和 **普通型** 变量进入 Test Pairs 下框中。

在 Test Type 框选择 Sign 方法。

4. 单击 OK 按钮, 提交运算。在输出窗口中, 得到输出结果, 见表 7-35、表 7-36。

表 7-35 Sign 检验频数表

Frequencies		N
普通型 - 豪华型减100	Negative ...	3
	Positive Differences ^a	5
	Ties ^c	2
	Total	10

a. 普通型 < 豪华型减100

b. 普通型 > 豪华型减100

c. 普通型 = 豪华型减100

表 7-36 Sign 检验统计量表

Test Statistics ^b	
	普通型 - 豪华型减100
Exact Sig. (2-tailed)	.727 ^a

a. Binomial distribution used.

b. Sign Test

5. 结果与讨论

表 7-35 列出了 **普通型** 小于 **豪华型减 100** 的观察值数量为 3, **普通型** 大于 **豪华型减 100** 的观察值数量为 5, 相等的有 2 个 (结)。

表 7-36 显示了符号检验双侧检验时, 用精确法得到的概率为 0.727。

本例为小样本, 故系统自动采用二项分布精确计算到的概率, 由于它大于 0.05, 故不拒绝 **普通型** 与 **豪华型减 100** 间无差异的原假设, 也即两种电器之间的差价满足厂家提出的相差 100 元的厂方建议。

7.7.3 麦内玛检验

在自身配对比较设计中，每个被试对象的响应分别在指定事件发生的前、后各被抽查 1 次。此时得到的数据是两分的，可以使用麦内玛（McNemar）检验。McNemar 检验确定初始的响应的比率（事件前）是否等于最终响应的比率（事件后）。它用来探究前、后设计中由于实验干涉引起响应的变化。

1. 表格构建

被研究的数据值限定为两个唯一响应类别。合计 $X_i < Y_i$ 的样品数 n_1 或 $X_i > Y_i$ 的样品数 n_2 。

2. 检验

如果 $n_1 + n_2 \leq 25$ ，当 $p = 0.5$ 和 $r = \min(n_1 + n_2)$ 时，在 $n_1 + n_2$ 次试验中， r 或少数“成功”事件的精确概率用下面的二项分布公式递推计算

$$p(X \leq r) = \sum_{i=0}^r \binom{n_1 + n_2}{i} (0.5)^{(n_1 + n_2)}$$

双侧检验的概率水平用计算得到的值加倍获得。如果 $n_1 + n_2 > 25$ ，则使用连续性修正的 χ^2 近似值

$$\chi_c^2 = \frac{(|n_1 - n_2| - 1)^2}{n_1 + n_2} \sim \chi^2(1)$$

当 $P(H_0) < \alpha$ 时，拒绝原假设。

3. 实例

例 7.19 某校某年级有 441 名学生。在统计学课程上，其中考试有 381 及格，60 人不及格，而期末考试中，有 367 人及格，74 人不及格。具体情况见表 7-37，表中数据已按要求存放在 data07-22.sav 的数据文件中，试用 McNemar 法检验期中考试和期末考试及格的比例是否相同？

表 7-37 统计学期中与期末考试合格情况表

		期末考试		合计
		及格	不及格	
期中考试	及格	329	52	381
	不及格	38	22	60
合计		367	74	441

由于期中考试成绩同期末考试成绩之间是相关的，又期中考试变量和期末考试变量

都是二元变量，只能在 0、1 中取一个值，故满足 McNemar 检验要求。具体步骤如下：

- (1) 在 SPSS 数据编辑窗口中，打开 data07-22.sav。
- (2) 加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法，用频数变量做加权处理。
- (3) 按 Analyze→Nonparametric Tests→Two-Related-Samples Tests 顺序，展开 Two-Related-Samples Tests 对话框。见图 7-10。

同时选中 *期中考试情况* 和 *期末考试情况* 变量进入 Test Pairs 下框中。

在 Test Type 框选择 McNemar 方法。

- (4) 单击 OK 按钮，提交运算。在输出窗口中，得到输出结果，见表 7-38、表 7-39。

- (5) 结果与讨论

表 7-38 显示了输入数据集中的基本情况的信息。它与表 7-37 中用黑框圈出的四个格中的数据相一致。

表 7-39 列出了 McNemar 检验中计算得到的各个统计量，从上到下各行依次为：总的观察值数量，即样本量为 441， χ^2 值为 1.878，在期中考试和期末考试及格的比例是相同的原假设下，出现目前统计量的值或者更极端值的双侧检验渐近分布计算得到的概率为 0.171，双侧检验用精确法计算得到的概率为 0.170，单侧检验用精确法计算得到的概率 0.085，点概率为 0.028。

表 7-38 McNemar 检验频数表

期中考试情况 & 期末考试情况		
期中考试情况	期末考试情况	
	不合格	合格
不合格	22	38
合格	52	329

表 7-39 McNemar 检验统计量表

Test Statistics ^b	
	期中考试情况 & 期末考试情况
N	441
Chi-Square ^a	1.878
Asymp. Sig.	.171
Exact Sig. (2-tailed)	.170
Exact Sig. (1-tailed)	.085
Point Probability	.028

a. Continuity Corrected

b. McNemar Test

由于本例为大样本，所以可用双侧检验渐近分布计算得到的概率 0.171 来进行判断，因它大于 0.05，故不拒绝原假设，认为期中考试和期末考试及格的比例是相同的。

7.7.4 边缘同质检验

如果数据是分类的，则使用边缘同质检验 (Marginal Homogeneity test)。它是 McNemar 检验从二项响应向多项式响应的扩展。它使用卡方分布检验实验干涉前、后设计中响应的变化。在 SPSS 中，只有安装了 ExactTest 选件时，边缘同质检验才有效。

例 7.20 按光洁度将产品分为三类：优等品、合格品和不合格品。两位检验员分别对 72 件产品进行检验，结果见表 7-40，表中数据已按要求存放在 data07-23.sav 的数据文件中，试用 Marginal Homogeneity test 法检验两位检验员的检验结果总体上是否一致？

表 7-40 两位检验员对某产品的检验结果表

检验员 1	检验员 2			合计
	优等	合格	不合格	
优等	17	4	8	29
合格	5	12	0	17
不合格	10	3	13	26
合计	32	19	21	72

由于两位检验员检验的是同一批产品，因此，他们给出的检验结果间是相关的。又这些检验结果是多项分类的，故满足 Marginal Homogeneity test 法的检验要求。具体步骤如下：

(1) 在 SPSS 数据编辑窗口中，打开 data07-23.sav。

(2) 加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法，用频数变量做加权处理。

(3) 按 Analyze→Nonparametric Tests→Two-Related-SamplesTests 顺序，展开 Two-Related-SamplesTests 对话框。见图 7-10。

同时选中 检验员 1 和 检验员 2 变量进入 Test Pairs 下框中。

在 Test Type 框选择 Marginal Homogeneity 方法。

(4) 单击 OK 按钮，提交运算。在输出窗口中，得到输出结果，见表 7-41。

(5) 结果与讨论

表 7-41 列出了 Marginal Homogeneity 检验中计算得到的各个统计量，从上到下各行依次为：分类数为 3，非对角线样品数为 30，观察到的 MH 统计量为 61，MH 统计量的均值为 57，MH 统计量的标准差为 4.583，标准 MH 统计量为 0.873，在两位检验员的检验结果总体上是一致的原假设下，出现目前统计量的值或者更极端值的双侧检验渐近分布计算得到的概率为 0.383。

由于 0.383 大于 0.05，故不拒绝原假设，而认为两位检验员的检验结果在总体上是一致的。

表 7-41 边缘同质检验结果

Marginal Homogeneity Test	
	检验员1 & 检验员2
Distinct Values	3
Off-Diagonal Cases	30
Observed MH Statistic	61.000
Mean MH Statistic	57.000
Std. Deviation of MH Statistic	4.583
Std. MH Statistic	.873
Asymp. Sig. (2-tailed)	.383

7.8 多个相关样本检验

当相关样本的数量达三个及三个以上时，需要用到本节介绍的方法。

7.8.1 弗里德曼检验

弗里德曼 (Friedman) 是等同于一个样本重复测定设计或每单元一个观察值的双因素方差分析的非参数检验。

1. 秩和的计算

对 N 个样品中的每一个样品， k 个变量被排序并被从 1 到 k 赋秩，在结上赋予平均秩。对 k 个变量中的每一个变量，计算样品的秩和。用符号 C_l 表示，则每一个变量的平均秩为： $R_l = C_l / N$ 。

2. 检验

此检验的无效假设： $H_0: k$ 个相关的变量来自同一个总体。

检验统计量为

$$\chi^2 = \frac{(12 / Nk(k+1)) \sum_{l=1}^k C_l^2 - 3N(k+1)}{1 - \sum T / Nk(k^2 - 1)}$$

其中， $\sum T = \sum_{i=1}^N \sum_{l=1}^k (t^3 - t)$ ， t 是变量结的长度。

在原假设为真时，上面的 $\chi^2 \sim \chi^2(k-1)$ 。

当 $P(H_0) < \alpha$ 时，拒绝原假设。

3. 实例

例 7.21 为比较 A、B、C、D 和 E 等 5 种药物注射后产生的皮肤疱疹的大小，选取 6 只家兔，给 6 只家兔先后分别按随机排列的次序注射这 5 种药物（每个家兔都注射 5 种药物），每种药物注射前，前一种药物所致疱疹痊愈，并对下一种药物致疱疹作用无影响。测试的实验结果（疱疹面积 mm^2 ）见表 7-42，表中数据已存放在 data07-24.sav 的数据文件中，这 5 种药物注射后产生的皮肤疱疹大小有差异吗？

表 7-42 注射 5 种药物测得的实验结果表

药物类别	家兔注射药物后产生的皮肤疱疹的面积					
	1	2	3	4	5	6
1	73	75	67	61	69	79
2	83	81	99	82	85	87
3	73	60	73	77	68	74
4	58	64	64	71	77	74
5	77	75	73	59	85	82

在 SPSS 中，解题步骤如下：

(1) 在 SPSS 数据编辑窗口中，打开 data07-24.sav。

(2) 按 Analyze→Nonparametric→K Related Samples 顺序展开 K Related Samples 对话框，见图 7-11。选择第 1 种药物实验结果，第 2 种药物实验结果，第 3 种药物实验结果，第 4 种药物实验结果和第 5 种药物实验结果变量进入 Test 框中。在 Test Type 框中选择 Friedman 方法。

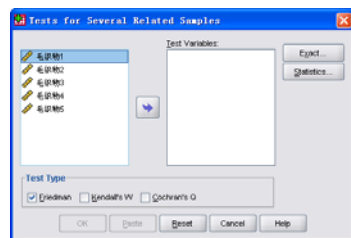


图 7-11 多个相关样本检验对话框

(3) 单击 OK 按钮，提交运算。在输出窗口中得到输出结果，见表 7-43 和表 7-44。

表 7-43 平均值表

Ranks	
	Mean Rank
第1种药物实验结果	2.50
第2种药物实验结果	4.92
第3种药物实验结果	2.25
第4种药物实验结果	1.92
第5种药物实验结果	3.42

表 7-44 检验统计量表

Test Statistics ^a	
N	6
Chi-Square	14.609
df	4
Asymp. Sig.	.006

a. Friedman Test

(4) 结果与讨论

表 7-43 列出了根据 5 种药物注射后产生的皮肤疱疹的面积，计算得到的平均秩。

表 7-44 列出了检验中的各个统计量，从上到下各行依次为：样本量为 6， $\chi^2 = 14.609$ ，自由度为 4，由渐近分布计算得到的概率为 0.006。

由于在 5 种药物注射后产生的皮肤疱疹大小无差异的原假设下，出现目前统计量的值或者更极端值的双侧检验的概率为 0.006，小于 0.05，故拒绝原假设，而认为这 5 种药物注射后产生的皮肤疱疹大小有差异。平均秩越小组产生的皮肤疱疹的面积越小。如果想进一步进行两两比较的检验，则应该使用前面提到的两个相关样本的威尔科克森检验法进行检验。

7.8.2 肯德尔调和系数 (Kendall's W) 检验

Kendall's W 是标准化的 Friedman 统计量, 是协(调)和系数。它是比率之间一致性的测度。每个样品是一个鉴定人或定价人, 每个变量是一个条件或被鉴定的人。对每个变量, 计算秩和。Kendall's W 范围在 0 (不同意) 和 1 (同意) 之间。对应于调查问卷中的多项排序题。

1. 协(调)和系数 W

$$W = \left(\frac{F}{N(k-1)} \right) \left(\frac{N^2 k(k^2 - 1)/12}{N^2 k(k^2 - 1)/12 - N \sum T/12} \right)$$

其中: F 是 Friedman 检验中的 χ^2 统计量, $\sum T = \sum_{i=1}^N \sum_{l=1}^k (t^3 - t)$, t 是变量结的长度。而 N 、 k 和 l 的含义同 Friedman 检验。

2. 检验

检验的无效假设: $H_0: \theta_1 = \theta_2 = \cdots = \theta_k$, 备择假设: $H_1: \theta_1, \theta_2, \cdots, \theta_k$ 不全相等。

在原假设为真时, $\chi^2 = N(k-1)W \sim \chi^2(k-1)$ 。

当 $P(H_0) < \alpha$ 时, 拒绝原假设。则认为各 θ_i 之间有一个顺序关系, 也即这 k 个观察值有这样的趋势: $x_{1j} \leq x_{2j} \leq \cdots \leq x_{kj}, x_{1j} < x_{kj}$, 这说明任意第 j 个区组内的 k 个观察值都有这样的趋势, 所以在 b 个区组中一致性趋于成立。

3. 实例

例 7.22 毛织物的一个重要的质量指标为织物的紧密程度。现有 9 位质检员对 5 种不同型号的毛织物进行手感评级, 每位质检员按各个毛织物的紧密程度用 1, 2, 3, 4 和 5 进行评定, 给出的数字越小, 表示该毛织物的紧密程度越大, 鉴定结果见表 7-45。表中数据已存放在 data07-25.sav 的数据文件中, 试问这 9 位质检员的评定结果是否一致?

表 7-45 9 位质检员的评定结果

		检验员								
		1	2	3	4	5	6	7	8	9
毛 织 物	1	2	2	2	1	1	1	1	1	2
	2	1	1	1	2	2	3	4	2	3
	3	3	3	3	3	4	4	3	5	1
	4	4	4	4	4	3	2	2	3	4
	5	5	5	5	5	5	5	5	4	5

在 SPSS 中, 解题步骤如下:

1. 在 SPSS 数据编辑窗口中, 打开 data07-25.sav。
2. 按 Analyze→Nonparametric→K Related Samples 顺序展开 K Related Samples 对话框, 见图 2-135。选择 *毛织物1*, *毛织物2*, *毛织物3*, *毛织物4* 和 *毛织物5* 变量进入 Test 框中。在 Test Type 框中选择 Kendall's W 方法。
3. 单击 OK 按钮, 提交运算。在输出窗口中得到输出结果, 见表 7-46 和表 7-47。

表 7-46 平均值表

Ranks	
	Mean Rank
毛织物1	1.44
毛织物2	2.11
毛织物3	3.22
毛织物4	3.33
毛织物5	4.89

表 7-47 检验统计量表

Test Statistics	
N	9
Kendall's694
Chi-Square	24.978
df	4
Asymp. Sig.	.000

a. Kendall's Coefficient of Concordance

4. 结果与讨论

表 7-46 列出了 9 名质检员对 5 种毛织物排名的平均秩。

表 7-47 列出了检验中的各个统计量, 从上到下各行依次为: 样本量为 9, Kendall's W 为 0.694, $\chi^2 = 24.978$, 自由度为 4, 由渐近分布计算得到的概率为 0.000。

由于在 5 种毛织物紧密程度相同的原假设下, 出现目前统计量的值或者更极端值的双侧检验的概率为 0.000, 小于 0.05, 故拒绝原假设, 认为 5 种毛织物的紧密程度是不同的, 正是这 9 位质检员对 5 种不同型号的毛织物进行手感评级时, 正确地区分了这 5 种不同型号的毛织物, 从而可认为这 9 位质检员的评定结果是一致的。

例 7.23 在对北京市 5 所体校运动员的问卷调查研究中, 186 名学生对“请在以下影响运动成绩的因素选项中, 按你认为的重要性的_{大小}做出排序选择: (1)运动强度 (2)运动量 (3)运动持续时间”做了回答, 数据存放在数据文件 data02-01.sav 中的*运动强度*、*运动量*、*运动持续时间*变量中。

SPSS 中的操作步骤如下:

- (1) 在 SPSS 数据编辑窗口中, 打开 data02-01.sav。
- (2) 按 Analyze→Nonparametric Tests→K Related Samples 顺序, 打开 Tests for Several Related Samples 对话框, 见图 7-11。在左侧变量名源框中, 选定*运动强度*、*运动量*、*运动持续时间*变量, 单击右移箭头, 将它们移入 Test Variables 框中。在 Test Type 框中选择 Kendall's W 方法。
- (3) 单击 OK 按钮运行, 在输出窗口中出现运算结果, 见表 7-48、表 7-49。
- (4) 结果与讨论

表 7-48 显示了*运动强度*的平均秩为 1.59, *运动量*的平均秩为 1.65, *运动持续时间*的平均秩为 2.77, 这同第 2 章中对例 2.52 用 Compute Variables 通过设置计算表达式计算得

到的结果是完全一致的,说明影响运动成绩因素中,第一位的是运动强度,第二位的是运动量,第三位的是运动持续时间。

表 7-49 则给出了 Kendall's W 检验的结果,样本量为 186, Kendall's W 为 0.444, $\chi^2 = 165.237$, 自由度为 2, 由渐近分布计算得到的概率为 0.000。

因在三个因素在重要性的排位上无差异的原假设下,出现目前统计量的值或者更极端值的双侧检验的概率为 0.000,故拒绝原假设,而认为三个因素在排位上是有差异的。

表 7-48 平均秩表

Ranks	
	Mean Rank
运动强度	1.59
运动量	1.65
运动持续时间	2.77

表 7-49 检验统计量

Test Statistics	
N	186
Kendall's444
Chi-Square	165.237
df	2
Asymp. Sig.	.000

a. Kendall's Coefficient of Concordance

(5) 结论

排位评定间差异有极显著性意义,5 所少体校的学生们对运动强度、运动量、运动持续时间三者在对运动成绩影响的重要性上有一致的看法,可以认为运动强度对运动成绩影响最重要,运动量次之,运动持续时间是第三位重要。

7.8.3 克科伦 Q 检验

克科伦 Q (Cochran's Q) 同 Friedman 检验是相同的,适用于所有的应答是两分的情况。它是对 k 个样本情形的 McNemar 检验的扩展。变量是在同一个个体或在配对个体上测定的。对应于调查问卷中的多选题。

1. 基本统计量的计算

对 N 个样品中的每一个样品,在 k 个指定的两分变量上取值,两分变量上第一个取到的值被当作“成功”处理,对每个样品合计“成功”变量的数量。样品 i “成功”的数量用 R_i 标记,变量 l 的总的“成功”的数量用 C_l 标记。

2. 检验

Cochran's Q 检验所作的原假设为: H_0 : 几个相关的两分变量有相同的均数。

Cochran's Q 用下式计算

$$Q = \frac{(k-1) \left[k \sum_{l=1}^k C_l^2 - \left(\sum_{l=1}^k C_l \right)^2 \right]}{k \sum_{l=1}^k C_l - \sum_{i=1}^N R_i^2}$$

在原假设为真时, $Q \sim \chi^2(k-1)$ 。

当 $P(H_0) < \alpha$ 时, 拒绝原假设。

3. 实例

例 7.24 186 名体校学生对问卷调查中的“如果你能在大学中继续从事你的运动专项, 在下面的选项中, 你最想上的地方是: (1) 大学高水平运动队, (2) 体育系, (3) 运动系, (4) 军警校, (5) 职业学院, (6) 其他。(可多选)”进行了回答, 结果存放在 data02-01.sav 数据文件的 Q5.1 至 Q5.6 中, 试用 Cochran's Q 法检验学生对这些去向的选择上是否有区别。

(1) 在 SPSS 数据编辑窗口中, 打开 data02-01.sav。

(2) 按 Analyze→Nonparametric→K Related Samples 顺序展开 K Related Samples 对话框, 见图 7-11。选择 Q5.1, Q5.2, Q5.3, Q5.4, Q5.5 和 Q5.6 变量进入 Test 框中。在 Test Type 框中选择 Cochran's Q 方法。

(3) 单击 OK 按钮, 提交运算。在输出窗口中得到输出结果, 见表 7-50 和表 7-51。

表 7-50 各项选择结果

	Frequencies	
	Value	
	0	1
上大学高水平运动队	106	80
上体育系	144	42
上运动系	73	113
上军警校	112	74
上职业学院	158	28
其他	180	6

表 7-51 检验统计量表

Test Statistics	
N	186
Cochran's Q	1.828E2 ^a
df	5
Asymp. Sig.	.000

a. 1 is treated as a success.

(4) 结果与讨论

表 7-50 列出了 6 项选择情况, 0 表示未选, 1 表示选择。

表 7-51 列出了检验中的各个统计量, 从上到下各行依次为: 样本量为 186, Cochran's Q 为 0.01828, 自由度为 5, 由渐近分布计算得到的概率为 0.000。

由于在学生毕业后去向的选择上没有区别的原假设下, 出现目前统计量的值或者更极端值的双侧检验的概率为 0.000, 它小于 0.05, 故拒绝原假设, 而认为学生对这些去向的选择上是有区别的。

7.9 交叉表资料的检验

在第 6 章和本章以上内容中, 介绍的差异和分布的一致性检验方法中, 虽然有一些也用以对定性数据的统计分析, 但总体上来看, 还是主要偏向于对实验研究中取得的定

量资料的统计分析的。所以本节将着重介绍一些对调查研究中收集到的一些数据资料（极大部分是定性资料）进行差异和关联性统计推断的方法。

一个行分类变量和一个列分类变量可以形成一个两维的交叉表格，再指定一个控制变量则可形成一个三维的交叉表。以此类推，可以形成一个多维的交叉表。许多统计书上还把这种表称作列联表。表中各个变量不同水平的交汇处，就是这种水平组合出现的频数或合计（Count）。

一般地，从某个总体 X 中随机抽取一个样本，按两个定性变量 A 、 B 分类，再将两个定性变量按属性或不同类型分为 r 组和 c 组，由此组成的资料称为 $r \times c$ 列联表。 $r \times c$ 列联表的基本形式见表 7-52。

表 7-52 $r \times c$ 列联表

		B						总计
		1	2	...	i	...	c	
A	1	f_{11}	f_{12}	...	f_{1j}	...	f_{1c}	r_1
	2	f_{21}	f_{22}	...	f_{2j}	...	f_{2c}	r_2
	\vdots	\vdots	\vdots	...	\vdots	...	\vdots	\vdots
	i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{ic}	r_i
	\vdots	\vdots	\vdots	...	\vdots	...	\vdots	\vdots
	r	f_{r1}	f_{r2}	...	f_{rj}	...	f_{rc}	r_r
总计		c_1	c_2	...	c_j	...	c_c	W

二维交叉表中的一个特殊的形式是行和列都取两个水平的情形，即属性 A 和属性 B 都是二分变量时，可以得到如下的表格，见表 7-53。由于调查研究中需要得到的核心内容只需表中黑框圈住的四个数，故称四格表。

表 7-53 四格表资料的形式

	B	\bar{B}	
A	f_{11}	f_{12}	r_1
\bar{A}	f_{21}	f_{22}	r_2
	c_1	c_2	W

当属性 A 和属性 B 不独立，且都取 2 个或 2 个以上的同等水平时，也会形成一个特殊的表格，见表 7-54，称这种表格为方表。

表 7-54 方表资料的形式

		化学考试				合计
		优秀	良好	及格	不及格	
物理考试	优秀	11	15	8	4	38
	良好	22	136	31	8	197
	及格	10	75	62	5	152
	不及格	5	9	7	6	27
合计		48	235	108	23	414

二维交叉表根据属性的类型，可以分为三类：属性 A 和属性 B 都为名义变量时的双向无序交叉表、属性 A 和属性 B 中有一个为名义变量，另一个为有序变量时的单向有序交叉表和两个都为有序变量时的双向有序交叉表。

7.9.1 二维交叉表资料的独立性检验

二维列联表的独立性检验问题实质上是分类数据的检验问题，主要讨论二维列联表的无方向检验问题。

7.9.1.1 皮尔逊卡方独立性检验

1. 检验

在属性 A 和属性 B 独立时，所要作的原假设为 $H_0: p_{ij} = \frac{f_{ij}}{r_i} = \frac{c_j}{W} \times \frac{r_i}{W}$ 。

式中符号代表的含义见表 7-52，即 f_{ij} 为变量 A 的第 i 个类别和变量 B 的第 j 个类别的观测频数， r_i 为变量 A 的第 i 个类别的观测总和， c_j 为变量 B 的第 j 个类别的观测总和， W 为总的观测次数。

在原假设为真时， $\chi_p^2 = \sum_{ij} \frac{(f_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((R-1)(C-1))$ 。其中 E_{ij} 为理论频数，

$E_{ij} = \frac{r_i c_j}{W}$ ， R 表示行水平数， C 表示列水平数。

当 $P(H_0) < \alpha$ 时，拒绝原假设。而认为 A 和 B 有关联。

本检验需要满足理论频数 $E_{ij} > 5$ 且 $df > 1$ 。

2. 实例

例 7.25 向 120 名女性和 120 名男性做调查，了解他们关于给谁买节日礼物最难的看法。调查结果见表 7-55。女性和男性在关于给谁买节日礼物最难的看法上有没有显著差异？

表 7-55 关于给谁买节日礼物最难的看法

性别	给谁买节日 物最难					
	配偶	父母	子女	兄弟姐妹	姻	其他 属
女性	28	34	23	7	13	15
男性	42	31	9	11	7	20

在 SPSS 中的解题步骤如下:

(1) 在 SPSS 数据窗口中, 将表 7-55 中的数据建成 SPSS 数据文件, 见 data07-26.sav。

(2) 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法, 用频数变量做加权处理。

(3) 单击 Analyze→Descriptive Statistics→Crosstabs, 展开 Crosstabs 对话框, 见图 2-90。在左侧源变量表中, 选择性别变量, 单击最上面的右移箭头将其移入 Row(s)框中, 选择送对象变量, 单击中间的右移箭头将其移入 Column(s)框中。

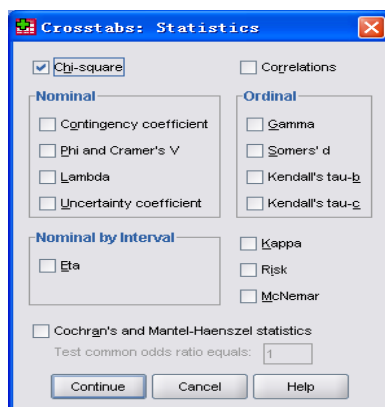


图 7-12 交叉表统计量选项卡

(4) 单击 Statistics 按钮, 打开 Statistics 对话框, 见图 7-12。选择 Chi-square 选项, 单击 Continue 按钮返回 Crosstabs 对话框。

(5) 单击 OK 按钮运行, 在输出窗口中得到输出表格, 见表 7-56、表 7-57 和表 7-58。

(6) 结果与讨论

表 7-56 列出了参与分析样品的基本情况, 包括有效样本量 240、有效百分比 100%, 缺失值数量 0、缺失值百分比 0%, 总样本量 240、百分比 100%。

表 7-57 列出了数据文件中的数据, 同表 7-55 中的数据相一致。

表 7-56 样品摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
性别 * 送礼对象	240	100.0%	0	.0%	240	100.0%

表 7-57 交叉表

性别 * 送礼对象 Crosstabulation							
性别		送礼对象					
		配偶	父母	子女	兄弟姐妹	姻亲	其他亲属
女	Count	28	34	23	7	13	15
男	Count	42	31	9	11	7	20
Total	Count	70	65	32	18	20	35

表 7-58 的备注 a.说明本例的期望频数没有出现小于 5 的情形, 最小期望频数为 9, 故使用皮尔逊 χ^2 检验是合适的。表中列出了 3 种检验方法的统计量、自由度和概率。最上面第一种检验方法即为皮尔逊 χ^2 检验。由于 $\chi^2 = 12.467$, $df = 5$, 在女性和男性在关于给谁买节日 物最难的看法没有差异的原假设下, 出现目前统计量的值或者更极端值

的双侧检验的概率为 0.029 小于 0.05, 故拒绝原假设, 而认为女性和男性在关于给谁买节日礼物最难的看法上有显著差异。

表 7-58 卡方检验

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	12.467 ^a	5	.029
Likelihood Ratio	12.736	5	.026
N of Valid Cases	240		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 9.00.

表中涉及的另一检验方法分别为: 似然比检验 (Likelihood Ratio)。将在下一节介绍。

在 SPSS 中, 进行多个独立总体多项分布的独立性卡方检验, 还可在 Analyze 的 Tables 过程中进行。

仍以例 7.25 为例, 在 SPSS 中具体操作步骤如下:

(1) 在 SPSS 数据窗口中, 打开数据文件 data07-26.sav。

(2) 按 Analyze→Tables→Custom Tables 顺序, 展开 Custom Tables 对话框 (见图 2-92)。

将性别拖曳到 Rows (行) 上, 将送对象拖曳到 Columns (列) 上, 注意: 性别和送对象必须定义为分类变量, 本例中它们被定义为名义测度。单击表格编辑器中的性别, 点亮 Define 下面的选择项, 单击 N% Summary Statistics 按钮, 展开 Summary statistics Categorical Variables 对话框, 见图 7-13。

在 Statistics 下框中, 选择 Row Total N%, 将它拖曳到 Display 下框的 Statistic 下。见图 7-13。单击 Apply to Selection 按钮, 返回到 Custom Tables 对话框 (见图 2-92)。

(3) 单击 Test Statistics 按钮, 展开 Test Statistics 对话框, 见图 7-14。选择卡方独立性检验 (Tests of independence(Chi-square))。

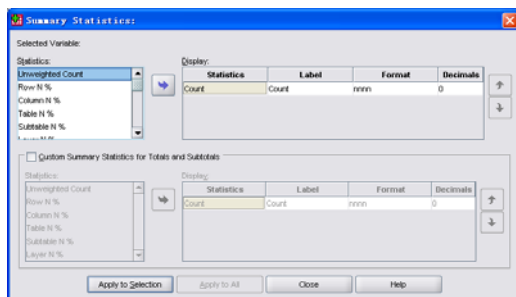


图 7-13 Summary statistics Categorical Variables 对话框

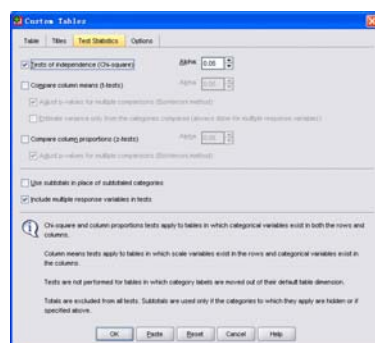


图 7-14 Test Statistics 对话框

单击确定按钮，则在输出窗口中得到计算结果，见表 7-59、表 7-60。

表 7-59 自定义表

Custom Table												
	送礼对象											
	1		2		3		4		5		6	
	Count	Row N %	Count	Row N %	Count	Row N %	Count	Row N %	Count	Row N %	Count	Row N %
性别 0	28	23.3%	34	28.3%	23	19.2%	7	5.8%	13	10.8%	15	12.5%
1	42	35.0%	31	25.8%	9	7.5%	11	9.2%	7	5.8%	20	16.7%

表 7-60 皮尔逊卡方检验表

Pearson Chi-Square Tests		
	送礼对象	
性别	Chi-square	12.467
	df	5
	Sig.	.029 ^a
Results are based on nonempty rows and columns in each innermost subtable.		
*. The Chi-square statistic is significant at the 0.05 level.		

(4) 结果与分析

表 7-59 中，显示了表 7-54 中的原始数据和对应的行百分比。

表 7-60 中，列出了皮尔逊卡方检验结果，卡方值为 12.467，自由度（df）为 5，在女性和男性在关于给谁买节日 物最难的看法没有差异的原假设下，出现目前统计量的值或者更极端值的双侧检验的概率为 0.029。故拒绝原假设，而认为女性和男性在关于给谁买节日 物最难的看法上有显著差异。

7.9.1.2 似然比独立性检验

1. 似然比检验（Likelihood Ratio）

在属性 A 和属性 B 独立时，所要作的原假设为 $H_0 : p_{ij} = \frac{f_{ij}}{r_i} = \frac{c_j}{W} \times \frac{r_i}{W}$ 。

式中符号代表的含义见表 7-52，即 f_{ij} 为变量 A 的第 i 个类别和变量 B 的第 j 个类别的观测频数， r_i 为变量 A 的第 i 个类别的观测总和， c_j 为变量 B 的第 j 个类别的观测总和， W 为总的观测次数。

在原假设为真时， $\chi^2_{LR} = -2 \sum_{ij} f_{ij} \ln \left(\frac{E_{ij}}{f_{ij}} \right) \sim \chi^2((R-1)(C-1))$ 。其中 E_{ij} 为理论频数，

$E_{ij} = \frac{r_i c_j}{W}$ ， R 表示行水平数， C 表示列水平数。

当 $P(H_0) < \alpha$ 时，拒绝原假设。而认为 A 和 B 有关联。

2. 实例

例 7.26 研究四个君子兰品种和某种病害发生的关系，得到的调查结果见表 7-61。试分析该病害的发生与品种是否有关？（天津农学院《园艺植物学科学研究导论》）

表 7-61 四个君子兰品种和某种病害发生关系的调查结果

品种感病情况	A	B	C	D	行合计
病株	45	30	15	27	117
健 株	155	170	185	173	683
列合计	200	200	200	200	800

在 SPSS 中的解题步骤如下：

(1) 在 SPSS 数据窗口中，将表 7-61 中的数据建成 SPSS 数据文件，见 data07-27.sav。

(2) 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法，用频数变量做加权处理。

(3) 按 Analyze→Descriptive Statistics→Crosstabs，展开 Crosstabs 对话框，见图 2-90。在左侧源变量表中，选择品种变量，单击最上面的右移箭头将其移入 Row(s): 框中，选择感病情况变量，单击中间的右移箭头将其移入 Column(s)框中。

(4) 单击 Statistics 按钮，打开 Statistics 对话框，见图 7-12。选择 Chi-square 选项[注：在 Crosstabs 中，似然比检验 (Likelihood Ratio) 没有单独列出，它隐含在 Chi-square 选项中]，单击 Continue 按钮返回 Crosstabs 对话框。

(5) 单击 OK 按钮运行，在输出窗口中得到输出表格，见表 7-62、表 7-63 和表 7-64。

(6) 结果与讨论

表 7-62 列出了参与分析样品的基本情况，包括有效样本量 800、有效百分比 100%，缺失值数量 0、缺失值百分比 0%，总样本量 800、百分比 100%。

表 7-63 列出了数据文件中的数据，同表 7-60 中的数据相一致。

表 7-62 样品摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
品种 * 感病情况	800	100.0%	0	.0%	800	100.0%

表 7-63 交叉表

品种 * 感病情况 Crosstabulation

Count		感病情况		Total
		病株	健壮株	
品种	A	45	155	200
	B	30	170	200
	C	15	185	200
	D	27	173	200
Total		117	683	800

表 7-64 似然比检验

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	18.290 ^a	3	.000
Likelihood Ratio	18.622	3	.000
N of Valid Cases	800		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 29.25.

从表 7-64 中第二行可见似然比检验 (Likelihood Ratio) 统计量值为 18.622，自由度为 3，在某病害的发生与品种无关的原假设下，出现目前统计量的值或者更极端值的双侧

检验的概率为 0.000, 由于它小于 0.05, 故拒绝原假设, 而认为该病害的发生与品种有关, 不同品种感染该病害有极显著性差异。

7.9.1.3 四格表资料独立性耶茨连续性修正检验法

1. 适用条件

在 7.9.1.1 中的 χ^2 检验, 适用于自由度大于 1 的情形, 所以要将它用于四格表资料的检验, 需对它进行修正, 这就要用到四格表 连续性修正 (Yates Continuity Corrected) 检验法。

连续性修正检验法适用于四格表大样本资料, 需要满足四个单元格中理论期望频数 $E_{ij} > 5$ 。

2. 检验

需要检验的原假设为 $H_0: p_1 = p_2$ 。

统计量为

$$\chi_c^2 = \begin{cases} \frac{W(|f_{11}f_{22} - f_{12}f_{21}| - 0.5W)^2}{r_1 r_2 c_1 c_2} & \text{如果 } |f_{11}f_{22} - f_{12}f_{21}| > 0.5W \\ 0 & \text{其他} \end{cases}$$

在原假设为真时, $\chi_c^2 \sim \chi^2(1)$ 。

当 $P(H_0) < \alpha$ 时, 拒绝原假设。而认为 A 和 B 有关联。

3. 实例

例 7.27 进行柱花草种子催 试验中, 调查经过 80℃ 热水 3~5 分钟种子与直接播种的种子的发 情况, 得到下表, 见表 7-65。试分析种子热水处理与否和种子 发多少是否有关?

表 7-65 柱花草种子发芽情况

处理项目	发 数	未发 数
种子处理	278	20
种子未处理	164	136

在 SPSS 中的解题步骤如下:

(1) 在 SPSS 数据窗口中, 将表 7-65 中的数据建成 SPSS 数据文件, 见 data07-28.sav。

(2) 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法, 用频数变量做加权处理。

(3) 按 Analyze→Descriptive Statistics→Crosstabs, 展开 Crosstabs 对话框, 见图 2-90。在左侧变量名源框中, 选择种子处理情况变量, 单击最上面的右移箭头将其移入 Row(s)

框中, 选择发 情况变量, 单击中间的右移箭头将其移入 Column(s)框中。

(4) 单击 Statistics 按钮, 打开 Statistics 对话框, 见图 7-12。选择 Chi-square 选项 (注: 在 Crosstabs 中, Yates Continuity Corrected for 2 x 2 Tables 没有单独列出, 它隐含在 Chi-square 选项中), 单击 Continue 按钮返回 Crosstabs 对话框。

(5) 单击 OK 按钮运行, 在输出窗口中得到输出表格, 见表 7-65、表 7-66 和表 7-67。

(6) 结果与讨论

表 7-66 列出了参与分析样品的基本情况, 包括有效样本量 598、有效百分比 100%, 缺失值数量 0、缺失值百分比 0%, 总样本量 598、百分比 100%。

表 7-67 列出了数据文件中的数据, 同表 7-65 中的数据相一致。

表 7-66 样品摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
种子处理情况 * 发芽情况	598	100.0%	0	.0%	598	100.0%

表 7-67 交叉表

种子处理情况 * 发芽情况 Crosstabulation			
Count		发芽情况	
		0	1
种子处理情况	0	136	164
	1	20	278
Total		156	442
		598	

表 7-68 耶茨连续性修正检验

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.157E2 ^a	1	.000	.000	.000
Continuity ...	113.659	1	.000		
Likelihood Ratio	126.509	1	.000		
Fisher's Exact Test					
N of Valid Cases ^b	598				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 77.74.

b. Computed only for a 2x2 table

从表 7-68 第二行中可见四格表 连续性修正检验统计量值为 113.659, 自由度为 1, 在种子热水处理与否和种子 发多少无关的原假设下, 出现目前统计量的值或者更极端值的双侧检验的概率为 0.000, 由于它小于 0.05, 故拒绝原假设, 而认为种子热水处理与否和种子 发多少有关, 热水处理能极显著地提高种子发 率。

本例同样可以使用 7.9.1.2 中介绍的似然比检验。似然比检验统计量值为 126.509, 自由度为 1, 双侧检验的概率为 0.000 (见表 7-65 第三行), 结论同 连续性修正检验结果一致。

7.9.1.4 四格表资料独立性费歇尔精确检验

1. 适用条件

如果四格表, 来源于单元格中不含缺失值的大表, 则当至少有一个期望单元格的合计数小于 5 时, 或者, 单纯四格表资料中, 至少有一个期望单元格的合计数小于 5 时, 需要用费歇尔精确检验 (Fisher's Exact Test)。

计算费歇尔精确检验的精确的单侧和双侧显著性水平, 适用于行和列独立假定及受观察边际总和约束下的四格表。全部单元格数被取整为最接近的整数。

2. 检验

需要检验的原假设为 $H_0: p_1 = p_2$

考察下面观察到的四格表:

n_1	n_2	$n_1 + n_2$
n_3	n_4	$n_3 + n_4$
$n_1 + n_3$	$n_2 + n_4$	N

受到观察边际总和的约束, 四个单元格数的值可以被表示为只有第一个单元格 n_1 的观察数。在独立的假设下, 第一个单元格中的期望频数 N_1 服从由下式给定 $N_1 = n_1$ 概率的超几何分布

$$P(N_1 = n_1) = \frac{(n_1 + n_2)!(n_3 + n_4)!(n_1 + n_3)!(n_2 + n_4)!}{N!n_1!n_2!n_3!n_4!}$$

其中, N_1 的范围从 $\max(0, n_1 - n_4)$ 到 $\min(n_1 + n_2, n_1 + n_3)$, 且 $N = n_1 + n_2 + n_3 + n_4$

精确单侧显著性水平 p_1 被定义为

$$p_1 = \begin{cases} P(N_1 \geq n_1) & \text{如果 } n_1 > E(N_1) \\ P(N_1 \leq n_1) & \text{如果 } n_1 \leq E(N_1) \end{cases}$$

其中, $E(N_1) = (n_1 + n_2)(n_1 + n_3) / N$

精确双侧显著性水平 p_2 被定义为单侧显著性水平 p_1 与 N_1 的样本空间另一侧所有点的概率的总和。使用 CDF.HYPER 累计分布函数计算单侧和双侧显著性水平。

当 $P(H_0) < \alpha$ 时, 拒绝原假设。而认为 A 和 B 有关联。

3. 实例

例 7.28 在对人脑的左右半球恶性和良性肿瘤的发病情况的调查中, 长在左半球的

12 个肿 中有 9 个良性, 3 个恶性, 而长在右半球的 4 个肿 中有 1 个良性, 3 个恶性, 请检验左、右半球的恶性肿 发病率是否有显著差异?

【题析】 由于本例样本量较少, 总共只有 16, 所以对此边际总和约束下的四格表资料的独立性检验, 可用费歇尔精确检验。

在 SPSS 中的解题步骤如下:

(1) 在 SPSS 数据窗口中, 将题中的数据建成 SPSS 数据文件, 见 data07-29.sav。

(2) 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法, 用频数变量做加权处理。

(3) 按 Analyze→Descriptive Statistics→Crosstabs, 展开 Crosstabs 对话框, 见图 2-90。在左侧变量名源框中, 选择半球分类变量, 单击最上面的右移箭头将其移入 Row(s):框中, 选择肿 分类变量, 单击中间的右移箭头将其移入 Column(s)框中。

(4) 单击 Statistics 按钮, 打开 Statistics 对话框, 见图 7-12。选择 Chi-square 选项(注: 在 Crosstabs 中, Fisher's Exact Test 没有单独列出, 它隐含在 Chi-square 选项中), 单击 Continue 按钮, 返回 Crosstabs 对话框。

(5) 单击 OK 按钮运行, 在输出窗口中得到输出表格, 见表 7-69、表 7-70 和表 7-71。

(6) 结果与讨论

表 7-69 列出了参与分析样品的基本情况, 包括有效样本量 16、有效百分比 100%, 缺失值数量 0、缺失值百分比 0%, 总样本量 16、百分比 100%。

表 7-70 列出了数据文件中的数据, 同题中数据相一致。

表 7-69 样品摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
半球分类 * 肿瘤分类	16	100.0%	0	.0%	16	100.0%

表 7-70 交叉表

半球分类 * 肿瘤分类 Crosstabulation

Count		肿瘤分类		Total
		恶性	良性	
半球分类	右半球	3	1	4
	左半球	3	9	12
Total		6	10	16

表 7-71 费歇尔精确检验

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.200 ^a	1	.074		
Continuity Correction ^b	1.422	1	.233		
Likelihood Ratio	3.175	1	.075		
Fisher's Exact Test				.118	.118
Linear-by-Linear Association	3.000	1	.083		
N of Valid Cases ^b	16				

a. 3 cells (75.0%) have expected count less than 5. The minimum expected count is 1.50.

b. Computed only for a 2x2 table

从表 7-71 第四行中可见, 在左、右半球的恶性肿瘤发病率没有差异的原假设下, 出现目前统计量的值或者更极端值的四格表费歇尔精确双侧检验的概率为 0.118, 由于它大于 0.05, 故现有证据不足以拒绝原假设, 故尚不能得出左、右半球的恶性肿瘤发病率有显著差异的统计结论。

表 7-70 第五行给出的是 Mantel-Haenszel 线性关联检验, 将在下文 7.9.1.6 中介绍。

7.9.1.5 四格表资料独立性的优比检验

1. 优比检验

优比检验又称为相对风险 (Relative Risk) 检验, 用 R 表示优比。它只适用于四格表 (即, $R=C=2$) 资料。

所做的原假设 $H_0: R=1$, 即属性 A 和属性 B 相互独立。

在病例对照研究, 相对风险用以下公式估计

$$R_0 = \frac{f_{11}f_{22}}{f_{12}f_{21}}$$

它被称为优比或相对风险。

相对风险 100 (1- α) 的百分位数置信区间用下面公式来获取

$$[R_0 \exp(-Z_{1-\alpha/2}v), R_0 \exp(Z_{1-\alpha/2}v)]$$

$$\text{其中, } v = \left(\frac{1}{f_{11}} + \frac{1}{f_{12}} + \frac{1}{f_{13}} + \frac{1}{f_{14}} \right)^{1/2}$$

在群组研究中两列的相对风险比都被计算。对于列 1, 风险是

$$R_1 = \frac{f_{11}(f_{21} + f_{22})}{f_{21}(f_{11} + f_{12})}$$

及对应的 100 (1- α) 的百分位数置信区间是

$$[R_1 \exp(-Z_{1-\alpha/2}v), R_1 \exp(Z_{1-\alpha/2}v)]$$

其中

$$v = \left(\frac{f_{12}}{f_{11}(f_{11} + f_{12})} + \frac{f_{22}}{f_{21}(f_{21} + f_{22})} \right)^{1/2}$$

对于列 2 的相对风险和置信区间的计算相似于列 1。

当 1 处在上述的置信区间之外, 表示原假设成立的概率小于 α , 从而拒绝原假设。

如果在有属性 A 的个体中有属性 B 的比例比没有属性 A 的个体中有属性 B 的比例高, 则优比 $R > 1$;

如果在有属性 A 的个体中有属性 B 的比例比没有属性 A 的个体中有属性 B 的比例低，则优比 $R < 1$ ；

如果属性 A 和属性 B 相互独立，则 $R = 1$ 。

2. 实例

例 7.29 为研究肺癌与吸烟之间是否存在关联，研究者在 30 至 35 的成年男子吸烟者中随机抽取了 68 名吸烟者的组成实验组，并在该年龄段的不吸烟者中随机抽取了 53 名不吸烟者组成对照组，20 年后，统计得到患肺癌患者的人数如表 7-72。

表 7-72 吸烟和不吸烟者患肺癌的人数统计表

	吸烟实验组	不吸烟对照组
有肺癌人数	62	6
无肺癌人数	37	16

(1) 在 SPSS 数据窗口中，将表 7-72 中的数据建成 SPSS 数据文件，见 data07-30.sav。

(2) 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法，用频数变量做加权处理。

(3) 按 Analyze→Descriptive Statistics→Crosstabs，展开 Crosstabs 对话框，见图 2-90。在左侧变量名源框中，选择组别变量，单击最上面的右移箭头将其移入 Row(s)框中，选择吸烟情况变量，单击中间的右移箭头将其移入 Column(s)框中。

(4) 单击 Statistics 按钮，打开 Statistics 对话框，见图 7-12。选择 Risk 选项，单击 Continue 按钮，返回 Crosstabs 对话框。

(5) 单击 OK 按钮运行，在输出窗口中得到输出表格，见表 7-73、表 7-74 和表 7-75。

表 7-73 样品摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
组别 * 吸烟情况	121	100.0%	0	.0%	121	100.0%

表 7-74 交叉表

组别 * 吸烟情况 Crosstabulation

Count		吸烟情况		Total
		不吸烟	吸烟	
组别	对照组	16	37	53
	肺癌组	6	62	68
Total		22	99	121

表 7-75 优比估计表

Risk Estimate			
	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for 组别 (对照组 / 肺癌组)	4.468	1.607	12.427
For cohort 吸烟情况 = 不吸烟	3.421	1.438	8.141
For cohort 吸烟情况 = 吸烟	.766	.632	.928
N of Valid Cases	121		

(6) 结果与讨论

表 7-73 列出了参与分析样品的基本情况, 包括有效样本量 121、有效百分比 100%, 缺失值数量 0、缺失值百分比 0%, 总样本量 121、百分比 100%。

表 7-74 列出了数据文件中的数据, 同表 7-72 中的数据相一致。

表 7-75 列出了优比统计量值、95%置信区间的下限和上限。

从上表可见, 三个区间中均不包括 1, 故拒绝 A 和 B 相互独立的原假设, 即拒绝肺癌与吸烟之间不存在关联的原假设。

(7) 结论

根据优比为 4.468 大于 1, 认为肺癌患者中吸烟的比例比对照组中吸烟的比例显著地高。

7.9.1.6 曼特-汉杰尔线性关联检验

1. 曼特-汉杰尔 (Mantel-Haenszel) 线性关联检验

在属性 A 和属性 B 独立时, 所要作的原假设为 $H_0: r = 0$, 即变量 A 和变量 B 之间线性无关。其中, r 是皮尔逊 (Pearson) 相关系数。

它需要两个变量为数量变量。

在原假设为真时, $\chi_{MH}^2 = (W - 1)r^2 \sim \chi^2(1)$ 。

式中, W 为总的观测次数。

当 $P(H_0) < \alpha$ 时, 拒绝原假设。而认为 A 和 B 有关联。

2. 实例

例 7.30 为研究导致死亡的结核病的类型和性别有没有关系, 调查了 3583 个因结核病死亡的人组成的样本, 并按性别和导致死亡的结核病的类型进行分类。数据见表 7-76。

表 7-76 结核病的类型和性别的关系

	男性	女性	合计
呼吸系统结核	2356	879	3235
其他类型结核	180	168	348
合计	2536	1047	3583

试分析导致死亡的结核病的类型和性别有没有关系?

在 SPSS 中的解题步骤如下:

(1) 在 SPSS 数据窗口中, 将题中的数据建成 SPSS 数据文件, 见 data07-31.sav。

(2) 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法, 用频数变量做加权处理。

(3) 按 Analyze→Descriptive Statistics→Crosstabs, 展开 Crosstabs 对话框, 见图 2-90。

在左侧变量名源框中，选择 **性别** 变量，单击最上面的右移箭头将其移入 Row(s) 框中，选择 **病因** 变量，单击中间的右移箭头将其移入 Column(s) 框中。

注：在数据文件中需将性别和病因变量的类型（Type）定义为数字型（Numeric）。

（4）单击 Statistics，打开 Statistics 对话框，见图 7-12。选择 Chi-square 选项（注：在 Crosstabs 中，Mantel-Haenszel 线性关联检验没有单独列出，它隐含在 Chi-square 选项中），单击 Continue 按钮返回 Crosstabs 对话框。

（5）单击 OK 按钮运行，在输出窗口中得到输出表格，见表 7-77、表 7-78 和表 7-79。

（6）结果与讨论

表 7-77 列出了参与分析样品的基本情况，包括有效样本量 3583、有效百分比 100%，缺失值数量 0、缺失值百分比 0%，总样本量 3583、百分比 100%。

表 7-78 列出了数据文件中的数据，同表 7-76 中的数据相一致。

表 7-77 样品摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
性别 * 病因	3583	100.0%	0	.0%	3583	100.0%

表 7-78 交叉表

性别 * 病因 Crosstabulation				
Count		病因		Total
		呼吸系统结核	其它类型结核	
性别	男	2356	180	2536
	女	879	168	1047
Total		3235	348	3583

表 7-79 Mantel-Haenszel 线性关联检验

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	67.662 ^a	1	.000		
Continuity Correction ^b	66.646	1	.000		
Likelihood Ratio	62.441	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	67.643	1	.000		
N of Valid Cases ^a	3583				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 101.69.

b. Computed only for a 2x2 table

从表 7-79 第五行中可见 Mantel-Haenszel 线性关联检验双侧检验的概率为 0.000，由于在性别和导致死亡的结核病的类型间线性无关的原假设下，出现目前统计量的值或者更极端值的概率为 0.000，小于 0.05，故拒绝原假设，认为导致死亡的结核病的类型和性别有关。

例 7.31 整理得到 278 例 体解剖资料见表 7-80，试问是否年龄越大的人，冠状动脉硬化的程度是否有越重的趋势？

在 SPSS 中的解题步骤如下：

（1）在 SPSS 数据窗口中，将表 7-80 中的数据建成 SPSS 数据文件，见 data07-32.sav。

（2）进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法，用频数变量做加权处理。

表 7-80 年龄与冠状动脉硬化的程度

年龄(岁)	冠状动脉硬化等级				合计
	1	2	3	4	
20~30	70	22	4	2	98
30~40	27	24	9	3	63
40~50	16	23	13	7	59
50 以上	9	20	15	14	58
合计	122	89	41	26	278

(3) 按 Analyze→Descriptive Statistics→Crosstabs, 展开 Crosstabs 对话框, 见图 2-90。在左侧变量名源框中, 选择年龄变量, 单击最上面的右移箭头将其移入 Row(s)框中, 选择冠状动脉硬化等级变量, 单击中间的右移箭头将其移入 Column(s)框中。

注: 在数据文件中需将年龄和冠状动脉硬化等级变量的类型 (Type) 定义为数字型 (Numeric)。

(4) 单击 Statistics 按钮, 打开 Statistics 对话框, 见图 7-12。选择 Chi-square 选项 (注: 在 Crosstabs 中, Mantel-Haenszel 线性关联检验没有单独列出, 它隐含在 Chi-square 选项中), 单击 Continue 按钮返回 Crosstabs 对话框。

(5) 单击 OK 按钮运行, 在输出窗口中得到输出表格, 见表 7-81、表 7-82 和表 7-83。

表 7-81 样品摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
年龄 * 冠状动脉硬化等级	278	100.0%	0	.0%	278	100.0%

表 7-82 交叉表

年龄 * 冠状动脉硬化等级 Crosstabulation						
Count		冠状动脉硬化等级				Total
		-	+	++	+++	
年龄	20至30	70	22	4	2	98
	30至40	27	24	9	3	63
	40至50	16	23	13	7	59
	50以上	9	20	15	14	58
Total		122	89	41	26	278

表 7-83 Mantel-Haenszel 线性关联检验

Chi-Square Tests				
	Value	df	Asymp. Sig. (2-sided)	
Pearson Chi-Square	71.432 ^a	9	.000	
Likelihood Ratio	73.739	9	.000	
Linear-by-Linear Association	63.389	1	.000	
N of Valid Cases	278			

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.42.

(6) 结果与讨论

表 7-81 列出了参与分析样品的基本情况, 包括有效样本量 278、有效百分比 100%,

缺失值数量 0、缺失值百分比 0%，总样本量 278、百分比 100%。

表 7-82 列出了数据文件中的数据，同表 7-80 中的数据相一致。

从表 7-83 第五行中可见，在年龄大小与冠状动脉硬化的程度线性无关的原假设下，出现目前统计量的值或者更极端值的概率为 0.000，小于 0.05，故拒绝原假设，而认为年龄越大的人，冠状动脉硬化的程度有越重的趋势。

7.9.1.7 方表的边缘齐性检验

麦内玛-克检验 (McNemar-Bowker's Test) 可以用来作属性 A 和属性 B 不独立时的对称方表 (包括四格表) 的边缘齐性检验。

1. 检验

给定一个 $n \times n$ 方表，McNemar-Bowker 统计量被用来检验假设 $H_0: p_{ij} = p_{ji} (i > j)$ ，备择假设 $H_1: p_{ij} \neq p_{ji}, (i, j)$ 中至少有一对不等。 p_{ij} 表示 i 行和 j 列的未知总体单元格概率。统计量用下述公式定义

$$\chi^2 = \sum_{i>j} \frac{I(n_{ij} + n_{ji} > 0)(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

其中， $I(\cdot)$ 函数是指示性函数。 n_{ij} 表示 i 行和 j 列的单元格中的观察数。

在原假设下， χ^2 服从自由度为 $n(n-1)/2$ 的渐近卡方分布。如果 $\chi^2 \geq \chi_{\alpha}^2(n(n-1)/2)$ ，

则拒绝原假设。双侧 p 值等于 $1 - CDF.CHISQ(\chi^2, n(n-1)/2)$ 。

2. 四格表特例

在四格表的属性 A 和属性 B 不独立时，对表 7-53 中，所做的原假设 $H_0: C_1 = R_1$ 的检验，就是边缘齐性检验。其中 C_1 、 R_1 分别表示第一列、第一行和的理论频数。

对于四格表，统计量缩减到能计算精确 p 值的经典 McNemar 统计量。双侧概率水平为

$$2 \sum_{i=0}^{\min(n_{12}, n_{21})} \binom{n_{12} + n_{21}}{i} (1/2)^{n_{12} + n_{21}}$$

当 $P(H_0) < \alpha$ 时，拒绝原假设。

3. 实例

例 7.32 张山和李四竞选某村村长，该村共有有效选民 557 个，竞选村长初期的底调查时有 362 人支持张山，195 人支持李四。在最后的选举时，有 372 人投票给了张山，

185 人把票投给了李四。557 个选民的交叉分组的情况见表 7-84。

表 7-84 557 个选民的交叉分组的情况表

		正式选举		合计
		支持张山	支持李四	
底调查	支持张山	334	28	362
	支持李四	38	157	195
合计		372	185	557

在这个结果中，我们要研究的是 底调查中支持张山后来却支持李四的比例和正式选举中本来支持李四最后却支持张山的比例是否相同，如果不相同，哪个比例大。

【题析】一般说来，在本类问题中，竞选初期 底调查的结果和正式选举的结果是相互关联的。底调查中支持张山的选民，一般在正式选举中也采取支持张山，反之亦然。

本例同四格表独立性卡方检验中遇到的问题看起来相似，实质上却有本质的区别。在本例中，同样是四格表，但表格中的属性 A 和 B 是不独立的。

我们所关心的 底调查中本来支持张山后来却支持李四的比例和 底调查中本来支持李四后来却支持张山的比例是否相同，同要问“底调查的分布情况与正式选举中分布情况是否一致？”是同一个问题。它属于概率四格表的边缘齐性检验问题。

在 SPSS 中的解题步骤如下：

- (1) 在 SPSS 数据窗口中，将表 7-83 中的数据建成 SPSS 数据文件，见 data07-33.sav。
- (2) 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法，用频数变量做加权处理。
- (3) 按 Analyze→Descriptive Statistics→Crosstabs，展开 Crosstabs 对话框，见图 2-90。在左侧变量名源框中，选择 底情况变量，单击最上面的右移箭头将其移入 Row(s)框中，选择 最后投票变量，单击中间的右移箭头将其移入 Column(s)框中。
- (4) 单击 Statistics 按钮，打开 Statistics 对话框，见图 7-12。选择 McNemar 选项，单击 Continue 按钮返回 Crosstabs 对话框。
- (5) 单击 OK 按钮运行，在输出窗口中得到输出表格，见表 7-85、表 7-86 和表 7-87。

表 7-85 样品摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
	摸底情况 * 最后投票	557	100.0%	0	0%	557

表 7-86 交叉表

摸底情况 * 最后投票 Crosstabulation				
Count		最后投票		Total
		选择张山	选择李四	
摸底情况	选举张山	334	28	362
	选举李四	38	157	195
Total		372	185	557

(6) 结果与讨论

表 7-85 列出了参与分析样品的基本情况，包括有效样本量 557、有效百分比 100%，缺失值数量 0、缺失值百分比 0%，总样本量 557、百分比 100%。

表 7-86 列出了数据文件中的数据，同表 7-83 中的数据相一致。

表 7-87 列出了 McNemar-Bowker's Test 结果。因为本例为四格表，所以，只输出精确检验的结果。在 底调查中支持张山后来却支持李四的比例和正式选举中本来支持李四最后却支持张山的比例是相同的原假设下，出现目前统计量的值或者更极端值的 McNemar-Bowker's Test 的双侧精确检验的概率为 0.268，大于 0.05，不拒绝原假设。

表 7-87 McNemar-Bowker's Test 表
Chi-Square Tests

	Value	Exact Sig. (2-sided)
McNemar Test		.268 ^a
N of Valid Cases	557	

a. Binomial distribution used.

(7) 结论

可以认为 底调查的分布情况与正式选举中分布情况是一致的。

例 7.33 两个中医对同一批 58 个患者的诊断结果见表 7-88，试问这两位中医的各项上的诊断正确率是否一致？

表 7-88 两个中医对同一批 58 个患者的诊断结果

		B 医生		
		阳虚	阴虚	阴阳两虚
A 医生	阳虚	25	3	1
	阴虚	1	9	1
	阴阳两虚	1	2	15

在 SPSS 中的解题步骤如下：

- (1) 在 SPSS 数据窗口中，将表 7-88 中的数据建成 SPSS 数据文件，见 data07-34.sav。
- (2) 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法，用频数变量做加权处理。
- (3) 按 Analyze→Descriptive Statistics→Crosstabs，展开 Crosstabs 对话框，见图 2-90。在左侧变量名源框中，选择 A 医生变量，单击最上面的右移箭头将其移入 Row(s)框中，选择 B 医生变量，单击中间的右移箭头将其移入 Column(s)框中。
- (4) 单击 Statistics 按钮，打开 Statistics 对话框，见图 7-12。选择 McNemar 选项，单击 Continue 按钮返回 Crosstabs 对话框。
- (5) 单击 OK 按钮运行，在输出窗口中得到输出表格，见表 7-89、表 7-90 和表 7-91。
- (6) 结果与讨论

表 7-89 列出了参与分析样品的基本情况，包括有效样本量 58、有效百分比 100%，缺失值数量 0、缺失值百分比 0%，总样本量 58、百分比 100%。

表 7-89 样品摘要

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
A医生 * B医生	58	100.0%	0	.0%	58	100.0%

表 7-90 交叉表

A医生 * B医生 Crosstabulation

	B医生			Total
	阳虚	阴虚	阴阳两虚	
A医生 阳虚	25	3	1	29
阴虚	1	9	1	11
阴阳两虚	1	2	15	18
Total	27	14	17	58

表 7-90 列出了数据文件中的数据，同表 7-87 中的数据相一致。

表 7-91 列出了 McNemar-Bowker's Test 结果。在这两位中医的各项上的诊断正确率是一致的原假设下，出现目前统计量的值或者更极端值的双侧渐近分布检验的概率为 0.721，大于 0.05，故不拒绝原假设。

表 7-91 McNemar-Bowker's Test 表

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
McNemar-Bowker Test	1.333	3	.721
N of Valid Cases	58		

(7) 结论

可以认为这两位中医的各项上的诊断正确率是一致的。

7.9.1.8 方表的一致性检验

科 (Cohen) 卡帕 (κ) 系数，它只对方表 ($R=C$) 定义，可用来检验方表的一致性检验。

1. 科 卡帕 (κ) 系数的计算公式为

$$\kappa = \frac{W \sum_{i=1}^R f_{ii} - \sum_{i=1}^R r_i c_i}{W^2 - \sum_{i=1}^R r_i c_i}$$

方差为

$$VAR_1 = W \left\{ \frac{\sum f_{ii} (W - \sum f_{ii})}{(W^2 - \sum r_i c_i)^2} + \frac{2(W - \sum f_{ii}) (2 \sum f_{ii} \sum r_i c_i - W \sum f_{ii} (r_i + c_i))}{(W^2 - \sum r_i c_i)^3} + \frac{(W - \sum f_{ii})^2 \left[W \sum_{i,j} f_{ij} (r_j + c_i)^2 - 4 \left(\sum r_i c_i \right)^2 \right]}{(W^2 - \sum r_i c_i)^4} \right\}$$

$$VAR_0 = \frac{1}{W \left(W^2 - \sum_i r_i c_i \right)^2} \left[W^2 \left(\sum_i r_i c_i \right) + \left(\sum_i r_i c_i \right)^2 - W \left(\sum_i r_i c_i (r_i + c_i) \right) \right]$$

式中符号代表的含义见表 7-52，即 f_{ii} 为变量 A 的第 i 个类别和变量 B 的第 i 个类别的观

测频数, r_i 为变量 A 的第 i 个类别的观测总和, c_i 为变量 B 的第 i 个类别的观测总和, W 为总的观测次数。

2. 检验

所要作的原假设为 $H_0: \kappa = 0$, 被择假设 $H_1: \kappa \neq 0$ 。

检验的统计量

$$U = \frac{\kappa}{\sqrt{VAR_0}} \xrightarrow{L} N(0,1)$$

当 $P(H_0) < \alpha$ 时, 拒绝原假设, 而认为 A 和 B 不是偶然一致的。

3. 应用中的注意点

科 的卡帕 (κ) 系数常用于医学和评价两种测量之间的一致性。其值在 -1 和 1 之间, 除根据 P 值判断一致性有无统计学意义外, 根据经验 $\kappa \geq 0.75$ 表明两者一致性较好, $0.75 \geq \kappa \geq 0.4$ 表明两者一致性一般, $\kappa < 0.4$ 表明两者一致性较差。

4. 实例

例 7.34 Glass1954 年对英国父 和儿子的社会地位的变化和流动情况做了一个抽样调查, 整理后的数据见表 7-92, 儿子的社会地位与父 的社会地位之间是否有一致性?

表 7-92 英国社会的变化和流动情况

		儿子的社会地位		
		上	中	下
父 的社会地位	上	588	395	159
	中	349	714	447
	下	114	320	441

在 SPSS 中的解题步骤如下:

(1) 在 SPSS 数据窗口中, 将表 7-92 中的数据建成 SPSS 数据文件, 见 data07-35.sav。

(2) 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法, 用频数变量做加权处理。

(3) 按 Analyze→Descriptive Statistics→Crosstabs, 展开 Crosstabs 对话框, 见图 2-90。在左侧变量名源框中, 选择父 的社会地位变量, 单击最上面的右移箭头将其移入 Row(s) 框中, 选择儿子的社会地位变量, 单击中间的右移箭头将其移入 Column(s) 框中。

(4) 单击 Statistics 按钮, 打开 Statistics 对话框, 见图 7-12。选择 kappa 选项, 单击 Continue 按钮返回 Crosstabs 对话框。

(5) 单击 OK 按钮运行, 在输出窗口中得到输出表格, 见表 7-93、表 7-94 和表 7-95。

表 7-93 样品摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
父亲的社会地位 * 儿子的社会地位	3527	100.0%	0	.0%	3527	100.0%

表 7-94 交叉表

父亲的社会地位 * 儿子的社会地位 Crosstabulation				
父亲的社会地位	儿子的社会地位	Count		
		上	中	下
上	中	588	395	159
中	中	349	714	447
下	中	114	320	441
Total		1051	1429	1047

(6) 结果与讨论

表 7-93 列出了参与分析样品的基本情况, 包括有效样本量 3527、有效百分比 100%, 缺失值数量 0、缺失值百分比 0%, 总样本量 3527、百分比 100%。

表 7-94 列出了数据文件中的数据, 同表 7-92 中的数据相一致。

表 7-95 列出了科 的 kappa 系数检验结果。科 的 kappa 系数值为 0.229(小于 0.4), 标准误差为 0.013, U 值为 19.201, 在儿子的社会地位与父 的社会地位之间没有一致性的原假设下, 出现目前统计量的值或者更极端值的双侧渐近分布检验的概率为 0.000, 由于它小于 0.05, 故拒绝原假设。

表 7-95 kappa 一致性检验表

Symmetric Measures				
		Value	Asymp. Std. Error ^a	Approx. T ^b
Measure of Agreement	Kappa	.229	.013	19.201
N of Valid Cases		3527		

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

(7) 结论

可以认为儿子的社会地位与父 的社会地位之间在统计学不是偶然一致的, 由于 kappa 系数值为 0.229, 故一致性较差。

7.9.1.9 二维交叉表的相合性检验

有序属性之间有两类相合关系: 正相合和负相合。

所谓正相合是指属性 A 比较大的个体属性 B 也往往比较大。而负相合是指属性 A 比较大的个体属性 B 却往往比较小。

在度量相合关系时, 用来计算相关系数的方法有许多: 如皮尔逊 (Pearson) 矩相关系数、斯皮尔曼 (Spearman) 秩相关系数、肯德尔 (Kendall) τ 相关系数、伽马 (Gamma) 系数和萨默斯 (Somers) 系数等。其中斯皮尔曼 (Spearman) 秩相关系数和肯德尔 (Kendall) τ 相关系数在交叉表中是常用的。

1. 皮尔逊矩相关系数

(1) 定义

设有成对数据 $\left(x_1, y_1\right), \cdots, \left(x_n, y_n\right)$, 则皮尔逊矩相关系数 r 的计算公式如下

$$r = \frac{\text{cov}(X, Y)}{\sqrt{S(X)S(Y)}} \equiv \frac{S}{T}$$

其中

$$\text{cov}(X, Y) = \sum_{i,j} X_i Y_j f_{ij} - \left(\sum_{i=1}^R X_i r_i \right) \left(\sum_{j=1}^C Y_j c_j \right) / W$$

$$S(X) = \sum_{i=1}^R X_i^2 r_i - \left(\sum_{i=1}^R X_i r_i \right)^2 / W$$

及

$$S(Y) = \sum_{j=1}^C Y_j^2 c_j - \left(\sum_{j=1}^C Y_j c_j \right)^2 / W$$

r 的方差为

$$\text{var}_1 = \frac{1}{T^4} \sum_{i,j} f_{ij} \left\{ T(X_i - \bar{X})(Y_j - \bar{Y}) - \frac{S}{2T} \left[(X_i - \bar{X})^2 S(Y) + (Y_j - \bar{Y})^2 S(X) \right] \right\}^2$$

如果无效假设是真的

$$\text{var}_0 = \frac{\sum_{i,j} f_{ij} X_i^2 Y_j^2 - \left(\sum_{i,j} f_{ij} X_i Y_j \right)^2 / W}{\left(\sum_i r_i X_i^2 \right) \left(\sum_j c_j Y_j^2 \right)}$$

其中

$$\bar{X} = \sum_{i=1}^R X_i r_i / W$$

和

$$\bar{Y} = \sum_{j=1}^C Y_j c_j / W$$

上述公式中， X_i 为升序排列 $X_1 < X_2 < \cdots < X_R$ 的行变量的相异值； Y_j 为升序排列

$Y_1 < Y_2 < \cdots < Y_C$ 的行变量的相异值; f_{ij} 为在单元格 (i, j) 中样品的权重和, $c_j = \sum_{i=1}^R f_{ij}$,

第 j 列和; $r_i = \sum_{j=1}^C f_{ij}$, 第 i 行和; $W = \sum_{j=1}^C c_j = \sum_{i=1}^R r_i$, 它是总合计。 R 是

行变量的类别数, C 是列变量的类别数。

矩相关系数的取值与两个变量值的大小有关。

(2) 检验

在 $\rho = 0$ 的假设下, $t = \frac{r\sqrt{W-2}}{\sqrt{1-r^2}}$ 服从自由度 $W-2$ 的 t 分布。

当 $P(H_0) < \alpha$ 时, 拒绝原假设, 而认为 A 和 B 是显著相合的。

(3) 应用条件

矩相关系数要求两个成对变量的总体服从正态分布, 它是一种参数性检验。

(4) 实例

例 7.35 从长期参加某场馆羽毛球健身活动的中青年中随机抽取 119 名男子进行调查, 结果得到他们每月的收入和用于健身活动的支出方面的调查数据资料见表 7-96, 试问这 107 名男子月收入与月健身支出间是否有正相合, 即月收入越高、月健身支出是否也越高?

表 7-96 119 名男子月收入与月健身支出

月收入 (元)	每月用于健身活动的支出 (元)				
	50 以下	50-100	100-150	150-200	200-250
1000-3000	2	4	2	0	0
3000-5000	3	6	10	2	0
5000-7000	1	7	13	6	1
7000-9000	0	9	16	7	2
9000-11000	0	2	9	7	3
11000-20000	0	0	2	4	1

在 SPSS 中的解题步骤如下:

- ① 在 SPSS 数据窗口中, 将表 7-96 中的数据建成 SPSS 数据文件, 见 data07-36.sav。
- ② 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法, 用频数变量做加权处理。
- ③ 整理月收入 and 月健身支出两个变量的频数分布表并做成数据文件

按 Analyze→Descriptive Statistics→Frequencies 展开 Frequencies 对话框, 选择月收入 and 月健身支出变量进入到 Variable[s]框中, 单击 OK 按钮运行, 则在输出窗口中可得到这

两个变量的频数分布表，并用月收入的上限和月健身支出的上限及其组内频数分别做成数据文件，见 data07-37.sav、data07-38.sav。

④ 数据资料的正态性检验

对其两个连续型变量的正态性检验，可参照例 7.4 中介绍的方法，先用各组的组中值作为各组的代表值，计算样本的加权均值和标准差，见表 7-97。然后用它来计算各组的累计频率，再计算各组的组内频率，对组内出现较少理论期望频数的组进行相邻组合并，使其组内频数满足大于 5 的要求，计算结果见 data07-37.sav、data07-38.sav，然后用适宜自由度的卡方拟合分布法进行正态分布的一致性检验。

本例的检验结果为不拒绝月收入 and 月健身支出两个变量服从正态分布的原假设。故可用 Pearson 的矩相关系数进行交叉表的相合分析。

⑤ 按 Analyze→Descriptive Statistics→Crosstabs，展开 Crosstabs 对话框，见图 2-90。在左侧变量名源框中，选择月收入变量，单击最上面的右移箭头将其移入 Row(s)框中，选择月健身支出变量，单击中间的右移箭头将其移入 Column(s)框中。

⑥ 单击 Statistics 按钮，打开 Statistics 对话框，见图 7-12。选择 Correlations 选项 (Pearson 的矩相关系数在其计算结果中)，单击 Continue 按钮返回 Crosstabs 对话框。

⑦ 单击 OK 按钮运行，在输出窗口中得到输出表格，见表 7-98、表 7-99 和表 7-100。

表 7-98 样品摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
月收入代表值 * 月健身支出代表值	119	100.0%	0	.0%	119	100.0%

表 7-97 样本均值和标准差

Descriptive Statistics			
	N	Mean	Std. Deviation
月收入代表值	119	7.2143E3	3108.60712
月健身支出代表值	119	1.2647E2	50.76680
Valid N (listwise)	119		

表 7-99 交叉表

月收入代表值 * 月健身支出代表值 Crosstabulation								
Count		月健身支出代表值					Total	
		25	75	125	175	250		
月收入代表值	2000	2	4	2	0	0	8	
	4000	3	6	10	2	0	21	
	6000	1	7	13	6	1	28	
	8000	0	9	16	7	2	34	
	10000	0	2	9	7	3	21	
	15500	0	0	2	4	1	7	
Total		6	28	52	26	7	119	

⑧ 结果与讨论

表 7-98 列出了参与分析样品的基本情况，包括有效样本量 119、有效百分比 100%，缺失值数量 0、缺失值百分比 0%，总样本量 119、百分比 100%。

表 7-99 列出了数据文件中的数据，同表 7-95 中的数据相一致。

表 7-100 列出了皮尔逊矩相关系数检验

表 7-100 Pearson 的矩相关系数表

Symmetric Measures				
	Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Interval by Interval Pearson's R	.444	.085	5.365	.000 ^c
Ordinal by Ordinal Spearman Correlation	.440	.073	5.296	.000 ^c
N of Valid Cases	119			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

结果。皮尔逊矩相关系数值（第一行，第二行为斯皮尔曼秩相关系数将在下一节介绍）为 0.444，标准误差为 0.065， T 值为 5.365，在皮尔逊矩相关系数为 0 的原假设下，出现目前统计量的值或者更极端值的双侧渐近分布检验的概率为 0.000，小于 0.05，故拒绝原假设。

⑨ 结论

由于 T 为 5.365 大于 0，说明 r 大于 0 ($t = \frac{r\sqrt{W-2}}{\sqrt{1-r^2}}$)，因此，可以认为这 107 名男

子月收入与月健身支出间有正相合。月收入越多，月健身支出越多。

2. 斯皮尔曼秩相关

(1) 定义

斯皮尔曼秩相关系数 r_s 是用 X_i 秩得分 R_i 和 Y_j 的秩得分 C_j 来计算的。这些秩得分定义如下

$$R_i = \sum_{k < i} r_k + (r_i + 1)/2, \quad i = 1, 2, \dots, R$$

$$C_j = \sum_{h < j} c_h + (c_j + 1)/2, \quad j = 1, 2, \dots, C$$

r_s 的公式和其渐近方差可以用 R_i 和 C_j 分别替代皮尔逊公式的 X_i 和 Y_j 来获得。

秩相关系数的大小与变量取值的顺序有关，与变量的取值无关。

(2) 检验

在 $\rho = 0$ 的假设下， $t = \frac{r\sqrt{W-2}}{\sqrt{1-r^2}}$ 服从自由度 $W-2$ 的 t 分布。

当 $P(H_0) < \alpha$ 时，拒绝原假设，而认为 A 和 B 是显著相合的。

以上公式中未说明的符号的含义参见皮尔逊矩相关系计算公式中的说明。

(3) 应用条件

斯皮尔曼秩相关不要求两个成对变量的总体服从正态分布，它是一种非参数性检验。

(4) 实例

例 7.36 500 个精神病人按 症和自 倾向的轻重程度的分类数据见表 7-101。
试问 程度大的人是否自 的倾向也越重？

表 7-101 500 个精神患者按抑郁症和自杀倾向的轻重程度的分类数据表

	无	中等	严重
无自 倾向	195	93	34
有自 倾向	20	27	27
曾自 过	26	39	39

【题析】 本例是个双向有序列联表。它要研究的是两个有序序列之间是否存在要大同大的趋势，只与变量位置值有关，而与变量本身值无关，它不需要两个变量服从正态性的假定，故可用斯皮尔曼秩相关系数来进行检验。

在 SPSS 中的解题步骤如下：

- ① 在 SPSS 数据窗口中，将表 7-100 中的数据建成 SPSS 数据文件，见 data07-39.sav。
- ② 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法，用频数变量做加权处理。
- ③ 按 Analyze→Descriptive Statistics→Crosstabs，展开 Crosstabs 对话框，见图 2-90。在左侧变量名源框中，选择 程度变量，单击最上面的右移箭头将其移入 Row(s)框中，选择 自 倾向变量，单击中间的右移箭头将其移入 Column(s)框中。
- ④ 单击 Statistics 按钮，打开 Statistics 对话框，见图 7-12。选择 Correlations 选项 (Spearman 秩相关系数在其计算结果中)，单击 Continue 按钮返回 Crosstabs 对话框。
- ⑤ 单击 OK 按钮运行，在输出窗口中得到输出表格，见表 7-102、表 7-103 和表 7-104。

表 7-102 样品摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
抑郁程度 * 自杀倾向	500	100.0%	0	.0%	500	100.0%

表 7-103 交叉表

抑郁程度 * 自杀倾向 Crosstabulation

Count		自杀倾向			Total
		无	想	曾	
抑郁程度	无	195	20	26	241
	中等	93	27	39	159
	严重	34	27	39	100
Total		322	74	104	500

⑥ 结果与讨论

表 7-102 列出了参与分析样品的基本情况，包括有效样本量 500、有效百分比 100%，缺失值数量 0、缺失值百分比 0%，总样本量 500、百分比 100%。

表 7-103 列出了数据文件中的数据，同表 7-100 中的数据相一致。

表 7-104 Spearman 秩相关系数表

Symmetric Measures					
		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Interval by Interval	Pearson's R	.356	.042	8.512	.000 ^c
Ordinal by Ordinal	Spearman Correlation	.368	.041	8.826	.000 ^c
N of Valid Cases		500			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

表 7-104 列出了斯皮尔曼秩相关系数检验结果。斯皮尔曼秩相关系数值（第二行）为 0.368（大于 0），标准误差为 0.041， T 值为 8.826，在斯皮尔曼秩相关系数为 0 的原假设下，出现目前统计量的值或者更极端值的双侧渐近分布检验的概率为 0.000，小于 0.05，故拒绝原假设。同样的理由，由于 T 值为 8.826 大于 0，所以，斯皮尔曼秩相关系数 $r_s > 0$ ，因此，程度大的人有自 的倾向越重的极显著的正相合。

注：由于本例为两个有序变量间的相合检验问题，只与两个变量的位置有关，与变量值无关，故不能用第一行的 Pearson 的矩相关系数值来解释。

⑦ 结论

可以认为 程度大的人自 的倾向也越重。

3. 适用于有序变量的其他相关系数

为使以下给出的计算有序变量的各个相关系数公式变得更加简明，特对各公式中的通用部分先做出定义：

设

$$D_r = W^2 - \sum_{i=1}^R r_i^2$$

$$D_c = W^2 - \sum_{j=1}^C c_j^2$$

$$C_{ij} = \sum_{h<i} \sum_{k<j} f_{hk} + \sum_{h>i} \sum_{k>j} f_{hk}$$

$$D_{ij} = \sum_{h<i} \sum_{k>j} f_{hk} + \sum_{h>i} \sum_{k<j} f_{hk}$$

$$P = \sum_{i,j} f_{ij} C_{ij}$$

$$Q = \sum_{i,j} f_{ij} D_{ij}$$

注：以上给出的 P 和 Q 是“平常的” P （同号对的数量）和 Q （异号对的数量）的两倍。同样地， D_r 是“平常的” $P+Q+X_0$ （同号对的数量，异号对的数量，以及行变量是结对的数量的两倍）和 D_c 是“平常的” $P+Q+Y_0$ （同号对的数量，异号对的数量，以及列变量是结对的数量的两倍）。

（1）肯德尔 $\tau-b$ 系数

$$\tau_b = \frac{P-Q}{\sqrt{D_r D_c}}$$

τ_b 的值在 -1 和 1 之间。 τ_b 的值越接近于 1, 越倾向于 A 和 B 之间有正相关, 它们有同步上升的趋势, 反之, 越接近于 -1, 越倾向于 A 和 B 之间有负相关。

其标准误差为

$$ASE_1 = \frac{1}{(D_r D_c)} \sqrt{\sum_{i,j} f_{ij} \left(2\sqrt{D_r D_c} (C_{ij} - D_{ij}) + \tau_b v_{ij} \right)^2 - W^3 \tau_b^2 (D_r + D_c)^2}$$

其中

$$v_{ij} = r_i D_c + C_j D_r$$

在独立假设下, 标准误差为

$$ASE_0 = 2 \sqrt{\frac{\sum_{i,j} f_{ij} (C_{ij} - D_{ij})^2 - \frac{1}{W} (P-Q)^2}{D_r D_c}}$$

(2) 肯德尔 $\tau-c$ 系数

$$\tau_c = \frac{q(P-Q)}{W^2(q-1)}$$

τ_c 的值在 -1 和 1 之间。 τ_c 的值越接近于 1, 越倾向于 A 和 B 之间有正相关, 它们有同步上升的趋势, 反之, 越接近于 -1, 越倾向于 A 和 B 之间有负相关。

其标准误差为

$$ASE_1 = \frac{2q}{(q-1)W^2} \sqrt{\sum_{i,j} f_{ij} (C_{ij} - D_{ij})^2 - \frac{1}{W} (P-Q)^2}$$

或, 在独立假设下

$$ASE_0 = \frac{2q}{(q-1)W^2} \sqrt{\sum_{i,j} f_{ij} (C_{ij} - D_{ij})^2 - \frac{1}{W} (P-Q)^2}$$

其中, $q = \min\{R, C\}$ 。

(3) 伽马 (γ) 系数

$$\gamma = \frac{P-Q}{P+Q}$$

γ 的值在-1 和 1 之间。 γ 的值越接近于 1, 越倾向于 A 和 B 之间有正相关, 它们有同步上升的趋势, 反之, 越接近于-1, 越倾向于 A 和 B 之间有负相关。

$$\text{其标准误差为: } ASE_1 = \frac{4}{(P+Q)^2} \sqrt{\sum_{i,j} f_{ij} (QC_{ij} - PD_{ij})^2}$$

$$\text{或在独立假设下, } ASE_0 = \frac{2}{(P+Q)} \sqrt{\sum_{i,j} f_{ij} (C_{ij} - D_{ij})^2 - \frac{1}{W} (P-Q)^2}$$

(4) 萨默斯 d 系数

为从行类别中预测列类别, 选择萨默斯 d 系数。

萨默斯 d 系数常用于 $2 \times c$ (或 $r \times 2$) 交叉表。当用“性别”作为行或列变量时, 行或列属性可以是无序的。对于 $2 \times c$ 交叉表, 用 $d_{B/A}$ 度量列属性 B 依赖于行属性 A 的相合性, 对于 $r \times 2$ 交叉表, 用 $d_{A/B}$ 度量行属性 A 依赖于列属性 B 的相合性。

萨默斯 d 系数用行变量 A 作为自变量, 计算公式为

$$d_{B/A} = \frac{P-Q}{D_r}$$

其标准误差为

$$ASE_1 = \frac{2}{D_r^2} \sqrt{\sum_{i,j} f_{ij} \{D_r (C_{ij} - D_{ij}) - (P-Q)(W - R_i)\}^2}$$

或在独立假设下

$$ASE_0 = \frac{2}{D_r} \sqrt{\sum_{i,j} f_{ij} (C_{ij} - D_{ij})^2 - \frac{1}{W} (P-Q)^2}$$

通过转换 A 和 B 的角色, 可以获取 Somers d 系数用 A 作为因变量的公式。

Somers d 系数的对称描述为

$$d = \frac{(P-Q)}{\frac{1}{2}(D_c + D_r)}$$

其标准误差为

$$ASE_1 = \frac{2\sigma_{\tau_b}^2}{D_r + D_c} \sqrt{D_r D_c}$$

其中, $\sigma_{\tau_b}^2$ 是 Kendall τ_b 的方差。

$$ASE_0 = \frac{4}{(D_c + D_r)} \sqrt{\sum_{i,j} f_{ij} (C_{ij} - D_{ij})^2 - \frac{1}{W} (P - Q)^2}$$

d 系数的值在-1 和 1 之间。其值越接近于 1, 越倾向于认为正相合, 反之, 越接近于-1, 越倾向于 A 和 B 之间有负相关。

(5) 对上述四种系数相合性检验

$H_0: A$ 与 B 独立

$$U = z / \sigma(z)$$

其中 $z = P - Q$, $\sigma(z)$ 为 z 的标准误。

$$\begin{aligned} \sigma^2(z) &= [W(W-1)(2W+5) - \sum r_i(r_i-1)(2r_i+5) - \sum c_j(c_j-1)(2c_j+5)]/18 \\ &+ [\sum r_i(r_i-1)(r_i-2) [\sum c_j(c_j-1)(c_j-2)]]/[9W(W-1)(W-2)] \\ &+ [\sum r_i(r_i-1) [\sum c_j(c_j-1)]]/[2W(W-1)] \end{aligned}$$

$\sigma^2(z)$ 的近似计算公式为

$$\sigma^2(z) \approx \frac{(W^3 - \sum r_i^3)(W^3 - \sum c_j^3)}{9W^3}$$

在原假设为真时, U 的渐近分布为 $N(0, 1)$ 。令 $\chi^2 = U^2 = \frac{z^2}{\sigma^2(z)}$, 则在原假设为真时, χ^2 的渐近分布为 $\chi_{(1)}^2$ 。

当 $P(H_0) < \alpha$ 时, 拒绝原假设, 而认为 A 和 B 是显著相合的。所计算的系数大于 0,

为正相合，反之为负相合。

(6) 实例

例 7.37 随机调查 26 位女职员和 24 位男职员的年收入情况，见表 7-105，试问收入和性别有没有关系？

表 7-105 26 位女职员和 24 位男职员的年收入情况

工资	女职员	男职员
22500-25000	1	0
25000-27500	4	1
27500-30000	2	1
30000-32500	10	3
32500-35000	3	5
35000-37500	5	6
37500-40000	1	6
40000-42500	0	2
合计	26	24

【题析】 这是一个典型的用 Somers d 系数处理的类型，可以用性别作为行变量，用工资作为列变量，对于 $2 \times c$ 交叉表，用 $d_{B|A}$ 度量列属性 B 依赖于行属性 A 的相合性。

在 SPSS 中的解题步骤如下：

- ① 在 SPSS 数据窗口中，将表 7-105 中的数据建成 SPSS 数据文件，见 data07-40.sav。
- ② 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法，用频数变量做加权处理。
- ③ 按 Analyze→Descriptive Statistics→Crosstabs，展开 Crosstabs 对话框，见图 2-90。在左侧变量名源框中，选择 *性别* 变量，单击最上面的右移箭头将其移入 Row(s) 框中，选择 *年收入* 变量，单击中间的右移箭头将其移入 Column(s) 框中。
- ④ 单击 Statistics 按钮，打开 Statistics 选项卡，见图 7-12。选择 Somers'd 选项，单击 Continue 按钮返回 Crosstabs 对话框。
- ⑤ 单击 OK 按钮运行，在输出窗口中得到输出表格，见表 7-106、表 7-107 和表 7-108。

表 7-106 样品摘要

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
性别 * 年收入	50	100.0%	0	.0%	50	100.0%

表 7-107 交叉表

性别*年收入 Crosstabulation

Count		年收入							
		22500-25000	25000-27500	27500-30000	30000-32500	32500-35000	35000-37500	37500-40000	40000-42500
性别	女	1	4	2	10	3	5	1	0
	男	0	1	1	3	5	6	6	2
Total		1	5	3	13	8	11	7	2
		Total							
		50							

⑥ 结果与讨论

表 7-106 列出了参与分析样品的基本情况,包括有效样本量 50、有效百分比 100%,缺失值数量 0、缺失值百分比 0%,总样本量 50、百分比 100%。

表 7-107 列出了数据文件中的数据,同表 7-104 中的数据相一致。

表 7-108 列出了 Somers'd 系数检验结果。根据题析,本列为特殊题型,应选用第三行的结果, $d_{\text{年收入|性别}} = 0.530$,标准误为 0.130, U 值为 4.062,在收入和性别无关(独立)的原假设下,出现目前统计量的值或者更极端值的双侧渐近分布检验的概率为 0.000,小于 0.05,故拒绝原假设,而认为收入和性别有关,又由于 $d = 0.530 > 0$,而数据文件中设置女性=0,男性=1,可知男性收入高于女性。

⑦ 结论

可以认为收入和性别有关,男性比女性收入高。

例 7.38 为了了解某种药品对某种疾病的疗效是否与患者的年龄有关,共抽查了 300 名患者。将疗效分为“显著”、“一般”和“较差”,将年龄分为“儿童”、“中青年”、“老年”。得如下的分布情况,见表 7-109。试问疗效和年龄是否有关?

表 7-109 疗效和年龄的对应关系

效果	年龄		
	儿童	中青年	老年
显著	58	38	32
一般	28	44	45
较差	14	18	23

在 SPSS 中的解题步骤如下:

- ① 在 SPSS 数据窗口中,将表 7-109 中的数据建成 SPSS 数据文件,见 data07-41.sav。
- ② 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法,用频数变量做加权

表 7-108 Somers'd 系数表

Directional Measures			Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Ordinal by Ordinal	Somers' d	Symmetric	.400	.098	4.062	.000
		性别 Dependent	.322	.078	4.062	.000
		年收入 Dependent	.530	.130	4.062	.000

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

处理。

③ 按 Analyze→Descriptive Statistics→Crosstabs, 展开 Crosstabs 对话框, 见图 2-90。在左侧变量名源框中, 选择 *年龄* 变量, 单击最上面的右移箭头将其移入 Row(s) 框中, 选择 *效果* 变量, 单击中间的右移箭头将其移入 Column(s) 框中。

④ 单击 Statistics 按钮, 打开 Statistics 对话框, 见图 7-12。选择 Gamma、Kendall's tau-b、Kendall's tau-c 选项, 单击 Continue 按钮, 返回 Crosstabs 对话框。

⑤ 单击 OK 按钮运行, 在输出窗口中得到输出表格, 见表 7-110、表 7-111 和表 7-112。

表 7-110 样品摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
年龄 * 效果	300	100.0%	0	.0%	300	100.0%

表 7-111 交叉表

年龄 * 效果 Crosstabulation

Count		效果			Total
		显著	一般	较差	
年龄	儿童	58	28	14	100
	中青年	38	44	18	100
	老年	32	45	23	100
Total		128	117	55	300

⑥ 结果与讨论

表 7-110 列出了参与分析样品的基本情况, 包括有效样本量 300、有效百分比 100%, 缺失值数量 0、缺失值百分比 0%, 总样本量 300、百分比 100%。

表 7-111 列出了数据文件中的数据, 同表 7-108 中的数据相一致。

表 7-112 列出了 Gamma、Kendall's tau-b、Kendall's tau-c 系数检验结果。在疗效和年龄无关 (独立) 的原假设下, 出现目前统计量的值或者更极端值的双侧渐近分布检验的概率为 0.000, 小于 0.05, 故拒绝原假设。又 Gamma、Kendall's tau-b、Kendall's tau-c 系数都为正值, 因而疗效和年龄间有极显著的正相合。

表 7-112 有序变量三种方法的系数表

Symmetric Measures					
		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	.182	.051	3.541	.000
	Kendall's tau-c	.177	.050	3.541	.000
	Gamma	.275	.076	3.541	.000
N of Valid Cases		300			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

⑦ 结论

可以认为疗效和年龄有关, 年龄越小疗效越高。

4. 适用于名义变量的关联程度分析

二维交叉表的卡方检验, 有显著性意义, 只是告诉你行列不是独立而是有关联的, 但它不告诉你关联的强度。因此, 需要用对称性测度来确定这种强度的数量。

在两个变量均为名义变量的条件下, 可以用以下几个统计量来反映二维列联表行列变量间关联程度的高低。

(1) 列联系数 (Coefficient of Contingency)

$$\text{计算公式为: } CC = \left(\frac{\chi_p^2}{\chi_p^2 + W} \right)^{1/2}$$

列联系数的取值范围为: $0 \leq CC < 1$ 。如果 $CC = 0$, 表示行列变量之间没有关联。它越接近于 1, 表示行列变量之间的关联越强, 从计算公式中可以看出, CC 受到样本量的影响, 样本量越大, CC 越小。

(2) φ (Phi) 系数

对于 2×2 表以外的表, $\varphi = \sqrt{\frac{\chi_p^2}{W}}$ 。只有在 2×2 表中, φ 等于皮尔逊相关系数, 使得

φ 的符号同皮尔逊相关系数的符号匹配。

φ 系数的绝对值越接近于 1, 表示行列变量之间的关联越强。

(3) Cramér 的 V 系数

$$\text{计算公式为: } V = \left(\frac{\chi_p^2}{W(q-1)} \right)^{1/2}, \text{ 其中, } q = \min\{R, C\}。$$

其取值在 0 到 1 之间。它越接近于 1, 表示行列变量之间的关联越强。

(4) 上述方法的检验

对上述系数的进行的显著性检验方法同二维交叉表的 χ^2 检验。

(5) 实例

例 7.39 在例 7.26 中, 我们已经知道四个君子兰品种和某种病害发生之间存在显著性的关联。试问关联度大吗? (原题中数据已存放在 Data07-27.sav 中)

① 在 SPSS 数据窗口中, 打开数据文件 data07-27.sav。

② 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法, 用频数变量做加权处理。

③ 按 Analyze→Descriptive Statistics→Crosstabs, 展开 Crosstabs 对话框, 见图 2-90。在左侧变量名源框中, 选择品种变量, 单击最上面的右移箭头将其移入 Row(s)框中, 选择感病情况变量, 单击中间的右移箭头将其移入 Column(s)框中。

单击 Statistics 按钮, 打开 Statistics 对话框, 见图 7-12。选择 Coefficient of Contingency、Phi and Cramér's V 选项, 单击 Continue 按钮返回 Crosstabs 对话框。

④ 单击 OK 按钮运行, 在输出窗口中得到输出表格, 见表 7-62、表 7-63 和表 7-113。

⑤ 结果与讨论

表 7-62、表 7-63 的解释参见例 7.26。

表 7-113 列出了 Phi 系数、Cramér 的 V 系数和关联系数及其检验结果。在四个君子兰品种和某种病害发生之间是独立的原假设下，出现目前统计量的值或者更极端值的双侧渐近分布检验的概率为 0.000，小于 0.05，故拒绝行和列独立的原假设。认为行列之间的关联是真实的，而不是由于偶然引起的。

表 7-113 关联系数、 φ 和 Cramér 的 V 系数表

Symmetric Measures		Value	Approx. Sig.
Nominal by Nominal	Phi	.151	.000
	Cramer's V	.151	.000
	Contingency Coefficient	.150	.000
N of Valid Cases		800	

⑥ 结论

关联虽然不是由于偶然引起，但三个方法的关联系数值都小于 0.2，说明君子兰品种和某种病害发生之间的关联度也不是很强。

5. 适用于名义变量的在预测中误差减少比例的测度分析

当上面这些测度给出某些关联强度的感 时，一般地，他们不会用直 来解释，为使这种感 更加清晰，应查看一下方向性测度。

方向性测度确定当你用已知的列变量值预测行变量值时误差减少的数量，反之亦然。常用的方向性测度系数主要包括以下几个：

(1) λ (Lambda) 系数

设 f_{im} 和 f_{mj} 分别是 i 行和 j 列中的最大单元格频数。同样，设 r_m 为最大的行总和， c_m 为最大的列总和。定义 $\lambda_{Y/X}$ 为用 X 类的信息预测单个 Y 类别时可以被忽略的相对误差的比例。 λ 定义误差为样品的错分类，并且样品根据模式（最常见）类别分类。

$\lambda_{Y/X}$ 的计算公式为

$$\lambda_{Y/X} = \frac{\sum_{i=1}^R f_{im} - c_m}{W - c_m}$$

标准误差为

$$ASE_0 = \frac{\sqrt{\sum_{i=1}^R \sum_{j=1}^C f_{ij} (\delta_{ij} - \delta_j)^2 - (\sum_{i=1}^R f_{im} - c_m)^2 / W}}{W - c_m}$$

$$ASE_1 = \frac{\sqrt{\sum_{i=1}^R \sum_{j=1}^C f_{ij} (\delta_{ij} - \delta_j + \lambda \delta_j)^2 - W \lambda_{Y/X}}}{W - c_m}$$

其中, $\delta_{ij} = \begin{cases} 1 & \text{如果 } j \text{ 是 } f_{im} \text{ 的列下标} \\ 0 & \text{其他} \end{cases}$

$\delta_j = \begin{cases} 1 & \text{如果 } j \text{ 是 } c_m \text{ 的下标} \\ 0 & \text{其他} \end{cases}$

对于用 Y 预测 X 的 λ , $\lambda_{Y/X}$, 是通过上面公式中序列改变下标来获取。

用两个不对称的 λ 的均值来获取对称的 λ 。

$$\lambda = \frac{\sum_{i=1}^R f_{im} + \sum_{j=1}^C f_{mj} - c_m - r_m}{2W - r_m - c_m}$$

标准误差为

$$ASE_0 = \frac{\sqrt{\sum_{i=1}^R \sum_{j=1}^C f_{ij} (\delta_{ij}^r + \delta_{ij}^c - \delta_i^r - \delta_j^c)^2 - \left[\left(\sum_{i=1}^R f_{im} + \sum_{j=1}^C f_{mj} - c_m - r_m \right)^2 / W \right]}}{2W - r_m - c_m}$$

$$ASE_1 = \frac{\sqrt{\sum_{i=1}^R \sum_{j=1}^C f_{ij} (\delta_{ij}^r + \delta_{ij}^c - \delta_i^r - \delta_j^c + \lambda (\delta_i^r + \delta_j^c))^2 - 4W}}{2W - r_m - c_m}$$

其中,

$\delta_{ij}^r = \begin{cases} 1 & \text{如果 } i \text{ 是 } f_{mj} \text{ 的行下标} \\ 0 & \text{其他} \end{cases}$

$\delta_i^r = \begin{cases} 1 & \text{如果 } i \text{ 是 } r_m \text{ 的下标} \\ 0 & \text{其他} \end{cases}$

以及

$$\delta_{ij}^c = \begin{cases} 1 & \text{如果 } j \text{ 是 } f_{im} \text{ 的列下标} \\ 0 & \text{其他} \end{cases}$$

$$\delta_i^c = \begin{cases} 1 & \text{如果 } j \text{ 是 } c_m \text{ 的下标} \\ 0 & \text{其他} \end{cases}$$

当自变量的值用来预测应变量值时，关联的测度反映了误差降低的比例。其值为 1 意味着自变量完全预测因变量。其值为 0 意味着自变量无助于预测因变量。

(2) Goodman 和 Kruskal 的 τ 系数

τ 定义误差为样品的错分类，并且样品被分到 j 类的概率等于 j 类的观察频率。

Goodman 和 Kruskal 的 τ 的计算公式为

$$\tau_{Y/X} = \frac{W \sum_{i,j} (f_{ij}^2 / r_i) - \sum_{j=1}^C c_j^2}{W^2 - \sum_{j=1}^C c_j^2}$$

标准误差为

$$ASE_1 = \sqrt{\frac{4}{\delta^4} \sum_{i,j} f_{ij} \left\{ \left(v - \delta \right) \left(\frac{1}{r_i} \sum_{j=1}^C f_{ij} c_j - c_j \right) - W \delta \left(\frac{1}{r^2} \sum_{j=1}^C f_{ij}^2 - \frac{1}{r_i} f_{ij} \right) \right\}}$$

其中， $\delta = W^2 - \sum_{j=1}^C c_j^2$ 及 $v = W \sum_{i,j} f_{ij}^2 / r_i - \sum_{j=1}^C c_j^2$

$\tau_{X/Y}$ 和其标准误差可以通过 X 和 Y 的角色互换来获取。

由于

$$(W-1)(C-1)\tau_{Y/X} \sim \chi_{(R-1)(C-1)}^2$$

$$(W-1)(C-1)\tau_{X/Y} \sim \chi_{(R-1)(C-1)}^2$$

所以，显著性水平可根据卡方分布给出。

(3) 不确定系数 (Uncertainty Coefficient)

不确定性系数的关联测度指出当一个变量的值被用来预测另一个变量的值时误差降低的比例。

设 $U_{Y/X}$ 是根据 X 的信息, Y 可以被忽略的不确定性 () 降低的比例。其计算公式为

$$U_{Y/X} = \frac{U(X) + U(Y) - U(XY)}{U(Y)}$$

其中, $U(X) = -\sum_{i=1}^R \frac{r_i}{W} \ln\left(\frac{r_i}{W}\right)$ 及 $U(Y) = -\sum_{j=1}^C \frac{c_j}{W} \ln\left(\frac{c_j}{W}\right)$

$$U(XY) = -\sum_{i,j} \frac{f_{ij}}{W} \ln\left(\frac{f_{ij}}{W}\right), \text{适用于 } f_{ij} > 0。$$

渐近标准误差为

$$ASE_1 = \frac{1}{WU(Y)^2} \sqrt{\sum_{i,j} f_{ij} \left\{ U(Y) \ln\left(\frac{f_{ij}}{r_i}\right) + \left[U(X) - U(XY) \ln\left(\frac{c_j}{W}\right) \right] \right\}}$$

$$ASE_0 = \frac{\sqrt{P - W[U(X) + U(Y) - U(XY)]^2}}{[WU(Y)]}$$

其中, $P = \sum_{i,j} f_{ij} \ln\left(\frac{c_j r_i}{W f_{ij}}\right)^2$

$U_{X/Y}$ 的公式可以通过转换 X 和 Y 的角色获得。

两个不对称的不确定性系数的对称描述定义如下

$$U = \left[\frac{U(X) + U(Y) - U(XY)}{U(X) + U(Y)} \right]$$

渐近的标准误差为

$$ASE_1 = \frac{2}{W[U(X) + U(Y)]^2} \sqrt{\sum_{i,j} f_{ij} \left\{ U(XY) \ln\left(\frac{r_i c_j}{W^2}\right) - [U(X) + U(Y)] \ln\left(\frac{f_{ij}}{W}\right) \right\}^2}$$

或

$$ASE_0 = \frac{2}{W[U(X) + U(Y)]} \sqrt{P - [U(X) + U(Y) - U(XY)]^2 / W}$$

(4) 实例

例 7.40 在例 7.26 中, 我们已经知道四个君子兰品种和某种病害发生之间存在显著性的关联。在例 7.39 中也对其行列的关联度进行了检验, 显然关联度不高, 如果用君子兰品种作为自变量, 用某种病害发生作为因变量, 则预测效果如何? (原题中数据已存放在 Data07-27.sav 中)

① 在 SPSS 数据窗口中, 打开数据文件 data07-27.sav。

② 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法, 用频数变量做加权处理。

③ 按 Analyze→Descriptive Statistics→Crosstabs, 展开 Crosstabs 对话框, 见图 2-90。在左侧源变量名表中, 选择品种变量, 单击最上面的右移箭头将其移入 Row(s)框中, 选择感病情况变量, 单击中间的右移箭头将其移入 Column(s)框中。

单击 Statistics 按钮, 打开 Statistics 对话框, 见图 7-12。选择 Lambda、Uncertainty Coefficient 选项, 单击 Continue 按钮返回 Crosstabs 对话框。

④ 单击 OK 按钮运行, 在输出窗口中得到输出表格, 见表 7-62、表 7-63 和表 7-114。

⑤ 结果与讨论

表 7-62、表 7-63 的解释参见例 7.26。

表 7-114 列出 λ 系数、Goodman 和 Kruskal 的 τ 系数和不确定系数及其检验结果。

由于在用品种变量的信息预测感病情况变量时可以被忽略的相对误差的比例等于 0 的原假设下, 对应于 λ 系数显著性检验出现目前统计量的值或者更极端值的概率为 0.103, 大于 0.05, 即 λ 系数检验结果表明, 用已知的品种变量值预测感病情况变量值时对误差减少的数量不显著。

表 7-114 λ 系数、Goodman 和 Kruskal 的 τ 系数和不确定系数表

Directional Measures							
Nominal by Nominal	Lambda	Symmetric	Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.	
		品种 Dependent	.042	.025	1.630	.103	
		感病情况 Dependent	.050	.030	1.630	.103	
			.000	.000			
	Goodman and Kruskal tau	品种 Dependent	.008	.003		.000 ^c	
		感病情况 Dependent	.023	.010		.000 ^c	
	Uncertainty Coefficient	Symmetric	.013	.006	2.207	.000 ^c	
		品种 Dependent	.008	.004	2.207	.000 ^c	
		感病情况 Dependent	.028	.013	2.207	.000 ^c	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Cannot be computed because the asymptotic standard error equals zero.

d. Based on chi-square approximation.

e. Likelihood ratio chi-square probability.

而用 Goodman 和 Kruskal 的 τ 系数及不确定性系数检验的结果表明, 渐近分布检验的概率为 0.000, 小于 0.05, 因而用已知的品种变量值预测感病情况变量值时误差减少的数量有显著性意义。但其值均为 0.008, 接近于 0, 所以预测效果不是很好。

⑥ 结论

除君子兰品种和某种病害发生之间的关联不是很强外, 用已知的品种变量值预测感病情况变量值时误差减少的数量很小, 效果不是很好。

6. 用 Eta 度量关联度

(1) 概述

当一个变量为分类变量, 另一个为数值变量时, 使用 Eta 来度量两者的关联度。分类

变量必须用数字编码。Eta 的平方表示由组间差异所解释的因变量方差的比例。

用列变量 Y 作因变量时的不对称 η 为

$$\eta_Y = \sqrt{1 - \frac{S_{YW}}{S(Y)}}$$

其中

$$S_{YW} = \sum_{i,j} Y_j^2 f_{ij} - \sum_{i=1}^R \frac{1}{r_i} \left(\sum_{j=1}^C Y_j f_{ij} \right)^2$$

Eta 在 0 和 1 之间取值, 0 表示行列变量间没有关联性, 接近于 1 表示行列变量间存在很高的关联性。

(2) 实例

例 7.41 调查得到某部门 50 名职工(女 26 名、男 24 名)的年收入情况, 见 data07-42.sav, 问性别和年收入之间的关联度怎样?

【题析】 用 7.5 中介绍的方法, 我们不难验证, 不同性别在年收入上是不同分布的, 男性收入高于女性, 有统计上的极显著性意义。

但两者间的关联程度有多高, 它们并没有给出, 这可以通过以下步骤来进行:

① 在 SPSS 数据窗口中, 打开数据文件 data07-42.sav。

② 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法, 用频数变量做加权处理。

③ 按 Analyze→Descriptive Statistics→Crosstabs, 展开 Crosstabs 对话框, 见图 2-90。在左侧源变量名表中, 选择性别变量, 单击最上面的右移箭头将其移入 Row(s)框中, 选择年收入变量, 单击中间的右移箭头将其移入 Column(s)框中。

单击 Statistics 按钮, 打开 Statistics 选项卡, 见图 7-12。在 Nominal by interval 下选择 Eta 选项, 单击 Continue 按钮返回 Crosstabs 对话框。

④ 单击 OK 按钮运行, 在输出窗口中得到三张输出表格, 第一张表为样品处理摘要, 第二张为计数表, 同上面介绍的许多例子中前两张输出表类似, 在此不再列出, 第三张为反映方向测度的 Eta 系数表, 见表 7-115。

⑤ 结果与讨论

表 7-115 列出了计算 Eta 系数的结果。第一行给出的是行变量为因变量时 Eta 系数的计算结果。第二行给出的是列变量为因变量时 Eta 系数的计算结果。显然本例应用第二行的结果来解

表 7-115 Eta 系数表

Directional Measures			Value
Nominal by Interval	Eta	性别 Dependent	.980
		年收入 Dependent	.508

释, 即性别和年收入之间的关联系数 Eta 为 0.508, 属中度关联。

⑥ 结论

该部门男性收入高于女性, 性别和年收入之间有中度关联。

7.9.2 多维交叉表资料的条件独立性和齐性检验

当对 n 个个体根据 3 个或 3 个以上的属性进行分类时, 将会得到三维或三维以上的交叉表, 即多维交叉表 (高维列联表)。由于多维交叉表的处理方式同三维交叉表的处理, 故这里只介绍对三维交叉表的处理。

1. 多维交叉表的压缩和分层

对多维交叉表资料的条件独立和齐性检验通常通过压缩和分层的方法, 将多维交叉表变成一个或多个二维交叉表, 再结合前面介绍的各种方法在各个层中分别对二维交叉表来进行检验。

(1) 压缩

在三个属性 A、B、C 的三维 $r \times c \times t$ 交叉表中, 合并其中一个属性的数据, 则得到另外两个属性的二维交叉表。

由于在 SPSS 中建立数据文件时, 一个属性对应于一个分类变量, 因此, 压缩的处理在 SPSS 中就相当于只选用未被合并的两个变量, 一个用来做行变量, 另一个用来做列变量, 此时得到的交叉表就是压缩一个变量后的数据资料。

对压缩后的二维交叉表的处理方法, 可根据两个变量的属性, 来分别选用上面介绍的相应方法。

(2) 分层

将多维交叉表按某一个属性分成几个低维交叉表是多维交叉表统计分析中又一个常用的方法, 称这种过程为多维交叉表的分层。

在 SPSS 中, 只需将分层变量依次选入到 Layer 框中, 即可完成分层工作。

2. 多维交叉表的条件独立性检验

在三个属性 A、B、C 的三维 $r \times c \times t$ 交叉表中, 需要考虑的条件独立性检验问题为: 原假设 H_0 : C 给定后 A 和 B 条件独立, 被择假设 H_1 : C 给定后 A 和 B 不条件独立。

按属性 C 来分层, 可将三维 $r \times c \times t$ 交叉表分成 t 个二维 $r \times c$ 交叉表。在原假设为真时, 其中每一个二维交叉表都是相互独立的。否则, 至少有一个二维 $r \times c$ 交叉表不是相互独立的。这就是说, 在分层二维交叉表独立性检验中, 只要有一个二维交叉表不独立, 原假设将被拒绝。反之, 当每一个二维交叉表都是相互独立时, 即可不拒绝原假设。

当要检验 A 给定后 B 和 C 条件独立或 B 给定后 A 和 C 条件独立的问题时, 其基本思路和上面是一样的。

例 7.42 某大学研究生院在当年 收新生中, 有 1659 名男生和 1413 名女生报考该

院,在五个专业中,共 收女生 422 名、男生 733 名,各专业的报考和 收的男、女生的详细情况见 data07-43.sav。试问各专业是否有偏爱男生的倾向?

【题析】 本例所要检验的原假设为专业给定后性别和录取情况间条件独立。故在 SPSS 中解题步骤如下:

① 在 SPSS 数据窗口中,打开数据文件 data07-43.sav。

② 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法,用频数变量做加权处理。

③ 按 Analyze→Descriptive Statistics→Crosstabs,展开 Crosstabs 对话框,见图 2-90。在左侧变量名源框中,选择性别变量,按最上面的右移箭头将其移入 Row(s)框中,选择录取情况变量,单击中间的右移箭头将其移入 Column(s)框中,选择专业变量,单击最下面的右移箭头将其移入 Layer 1 of 1 下框中,把专业设定为第一层(即最里层)的层变量。

单击 Statistics 按钮,打开 Statistics 对话框,见图 7-12。选择 Chi-square 选项,单击 Continue 按钮返回 Crosstabs 对话框。

④ 单击 OK 按钮运行,在输出窗口中得到三张输出表格,见表 7-116、表 7-117、表 7-118。

⑤ 结果与讨论

表 7-116 样品处理摘要表,表 7-117 计数表,同上面介绍的许多例子中前两张输出表

表 7-116 样品处理摘要表

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
性别 * 录取情况 * 专业	3072	100.0 %	0	.0 %	3072	100.0 %

表 7-118 分层检验表

Chi-Square Tests						
专业		Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
A1	Pearson Chi-Square	.017 ^a	1	.695		
	Continuity Correction ^b	.000	1	1.000		
	Likelihood Ratio	.017	1	.895		
	Fisher's Exact Test				1.000	.514
	Linear-by-Linear Association	.017	1	.895		
	N of Valid Cases ^c	700				
A2	Pearson Chi-Square	.206 ^d	1	.650		
	Continuity Correction ^b	.150	1	.699		
	Likelihood Ratio	.206	1	.650		
	Fisher's Exact Test				.688	.349
	Linear-by-Linear Association	.206	1	.650		
	N of Valid Cases ^c	970				
A3	Pearson Chi-Square	.237 ^e	1	.627		
	Continuity Correction ^b	.168	1	.691		
	Likelihood Ratio	.236	1	.627		
	Fisher's Exact Test				.657	.345
	Linear-by-Linear Association	.236	1	.627		
	N of Valid Cases ^c	584				
A4	Pearson Chi-Square	.039 ^f	1	.844		
	Continuity Correction ^b	.008	1	.927		
	Likelihood Ratio	.039	1	.844		
	Fisher's Exact Test				.835	.461
	Linear-by-Linear Association	.039	1	.844		
	N of Valid Cases ^c	571				
A5	Pearson Chi-Square	.013 ^f	1	.910		
	Continuity Correction ^b	.000	1	1.000		
	Likelihood Ratio	.013	1	.911		
	Fisher's Exact Test				1.000	.548
	Linear-by-Linear Association	.013	1	.911		
	N of Valid Cases ^c	247				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 17.43.

b. Computed only for a 2x2 table.

c. 0 cells (.0%) have expected count less than 5. The minimum expected count is 158.61.

d. 0 cells (.0%) have expected count less than 5. The minimum expected count is 97.27.

e. 0 cells (.0%) have expected count less than 5. The minimum expected count is 44.06.

f. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.78.

表 7-117 分层计数表

性别 * 录取情况 * 专业 Crosstabulation				
		录取情况		Total
		未录取	录取	
A1 专业	性别 女	17	33	50
	男	227	423	650
	Total	244	456	700
A2 专业	性别 女	345	200	545
	男	263	162	425
	Total	608	362	970
A3 专业	性别 女	184	90	274
	男	214	96	310
	Total	398	186	584
A4 专业	性别 女	296	94	390
	男	136	45	181
	Total	432	139	571
A5 专业	性别 女	143	11	154
	男	96	7	93
	Total	229	18	247

类似, 在此不再解释。

表 7-118 列出了各层检验结果。以似然比检验结果为例, 在五个专业中, 似然比值均小于 0.3, 在各层独立的原假设下, 出现目前统计量的值或者更极端值的概率都大于 0.05, 因而, 总体而言, 不能拒绝原假设。所以认为条件独立成立。

⑥结论

各个专业没有偏爱男生的倾向。

3. 特定三维 $r \times 2 \times 2$ 交叉表的条件独立性和齐性 (同质性) 检验

在 SPSS 中, 专门提供了一种检验特定多维交叉表的克科伦和曼特-汉杰尔 (Cochran's and Mantel-Haenszel) 法来检验条件独立性和齐性检验。

(1) 条件独立性检验

Cochran's and Mantel-Haenszel 统计量, 它通过控制一个或多个其它分类变量, 用它们来定义层的数量, 从而实现在有条件约束下, 检验二分类因素变量和二分类响应变量之间的独立性。

原假设 H_0 : A 给定后 B 和 C 条件独立, 被择假设 H_1 : A 给定后 B 和 C 不条件独立。

Cochran 统计量为

$$C = \frac{\sum_{k=1}^K w_k d_k / \sum_{k=1}^K w_k}{\sqrt{\sum_{k=1}^K w_k \hat{p}_k (1 - \hat{p}_k) / \sum_{k=1}^K w_k}} = \frac{\sum_{k=1}^K w_k d_k}{\sqrt{\sum_{k=1}^K w_k \hat{p}_k (1 - \hat{p}_k)}}$$

Cochran 统计量还可以表述为

$$C = \frac{\sum_{k=1}^K (f_{11k} - E_{11k})}{\sqrt{\sum_{k=1}^K w_k \hat{p}_k (1 - \hat{p}_k)}}$$

其中

K 为层的数量。

f_{ijk} 为在第 k 层的第 j 列第 i 行样品的单元格权重和。

c_{jk} 等于 $\sum_{i=1}^R f_{ijk}$, 它是第 k 层第 j 列的小计。

r_{ik} 等于 $\sum_{j=1}^C f_{ijk}$, 它是第 k 层第 i 行的小计。

$n_k = \sum_{j=1}^C c_{jk} = \sum_{i=1}^R r_{ik}$, 它是第 k 层的合计。

E_{ijk} 为第 k 层的第 j 列第 i 行期望单元格计数, $E(f_{ijk}) = r_{ik} c_{jk} / n_k$ 。

而

$$\hat{p}_{ik} = \frac{f_{ik}}{r_{ik}}$$

$$d_k = \hat{p}_{1k} - \hat{p}_{2k}$$

$$\hat{p}_k = \frac{c_{1k}}{n_k}$$

以及

$$w_k = \frac{r_{1k} r_{2k}}{n_k}$$

当层的数量固定时, 由于各层内样本量的增加, Cochran 统计量近似服从标准正态, 因此, 其平方近似服从自由度为 1 的 χ^2 分布。

Mantel and Haenszel 的统计量只不过是样本小时对连续性和方差“膨胀”进行修正的 Cochran 统计量。当 r_{1k} 和 r_{2k} 较小时, 这些修正是可取的, 而且即使 r_{1k} 和 r_{2k} 比较大时, 修正可以获得显著差异。Mantel and Haenszel 统计量被定义为

$$M = \frac{\left\{ \sum_{k=1}^K (f_{11k} - E_{11k}) \right\} - 0.5 \left\{ \sum_{k=1}^K (f_{11k} - E_{11k}) \right\}}{\sqrt{\sum_{k=1}^K \frac{r_{1k} r_{2k}}{n_k - 1} \hat{p}_k (1 - \hat{p}_k)}}$$

其中, sgn 是正负号函数

$$\text{sgn}(x) = \begin{cases} 1, & \text{如果 } x > 0 \\ 0, & \text{如果 } x = 0 \\ -1, & \text{如果 } x < 0 \end{cases}$$

当层的数量固定时, 由于各层内样本量增加, 或当每层内的样本量被固定, 由于层的数量的增加, 该统计量近似标准正态, 因此, 其平方近似服从自由度为 1 的 χ^2 分布。

(2) 齐性检验

Breslow-Day 统计量被用来对公共优比 (common odds ratio) 进行齐性检验, 它是用 Cochran 和 Mantel-Haenszel 统计量检验的弱势条件同条件独立之比 (也就是, 齐性时公

共优比为 1)。

齐性检验时,所要检验的原假设 H_0 : 优比(OR)为 1。对任何一个估计量, Breslow-Day

统计量 $\hat{\theta}$ 为: $\sum_{k=1}^K \frac{\{f_{11k} - E(f_{11k} | c_{1k}; \hat{\theta})\}^2}{V(f_{11k} | c_{1k}; \hat{\theta})}$ 。E 和 V 是基于严格的矩,但习惯上用渐近期望

和方差来替代它们。设 \hat{E} 和 \hat{V} 分别表示估计的渐近期望和估计的渐近方差。给出 Mantel-Haenszel 公共优比估计量 $\hat{\theta}_{MH}$, 我们使用下面的统计量作为 Breslow-Day 统计量

$$B = \sum_{k=1}^K \frac{\{f_{11k} - \hat{E}(f_{11k} | c_{1k}; \hat{\theta}_{MH})\}^2}{\hat{V}(f_{11k} | c_{1k}; \hat{\theta}_{MH})}$$

其中, $\hat{E}(f_{11k} | c_{1k}; \hat{\theta}_{MH}) = \hat{f}_{11k}$, 满足等式 $\frac{\hat{f}_{11k}(n_k - r_{1k} - c_{1k} + \hat{f}_{11k})}{(r_{1k} - \hat{f}_{11k})(c_{1k} - \hat{f}_{11k})} = \hat{\theta}_{MH}$, 有约束条件使得:

$\hat{f}_{11k} \geq 0$, $(r_{1k} - \hat{f}_{11k}) > 0$, $(c_{1k} - \hat{f}_{11k}) > 0$, $n_k - r_{1k} - c_{1k} + \hat{f}_{11k} \geq 0$, 并且

$\hat{V}(f_{11k} | c_{1k}; \hat{\theta}_{MH}) = \left(\frac{1}{f_{11k}} + \frac{1}{f_{12k}} + \frac{1}{f_{21k}} + \frac{1}{f_{22k}} \right)^{-1}$, 有约束条件使得: $\hat{f}_{11k} > 0$,

$\hat{f}_{12k} = (r_{1k} - \hat{f}_{11k}) > 0$, $\hat{f}_{21k} = (c_{1k} - \hat{f}_{11k}) > 0$, $\hat{f}_{22k} = n_k - r_{1k} - c_{1k} + \hat{f}_{11k} > 0$ 。

在优比为常量的无效假设下, Breslow-Day 统计量近似服从自由度为 $K-1$ 的 χ^2 随机变量渐近分布。

Tarone 统计量是用于一致性检验的 Breslow-Day 统计量的修正。

当公共优比估计量是一致但效率低时,尤其是当我们得到 Mantel-Haenszel 公共优比估计量时, Tarone (1985) 建议校正 Breslow-Day 统计量。对于 $\hat{\theta}_{MH}$, Tarone 的校正统计量为

$$T = \sum_{k=1}^K \frac{\{f_{11k} - \hat{E}(f_{11k} | c_{1k}; \hat{\theta}_{MH})\}^2}{\hat{V}(f_{11k} | c_{1k}; \hat{\theta}_{MH})} - \frac{\left[\sum_{k=1}^K \{f_{11k} - \hat{E}(f_{11k} | c_{1k}; \hat{\theta}_{MH})\} \right]^2}{\sum_{k=1}^K \hat{V}(f_{11k} | c_{1k}; \hat{\theta}_{MH})}$$

$$= B - \frac{\left[\sum_{k=1}^K \left\{ f_{11k} - \hat{E}(f_{11k} | c_{1k}; \hat{\theta}_{MH}) \right\} \right]^2}{\sum_{k=1}^K \hat{V}(f_{11k} | c_{1k}; \hat{\theta}_{MH})}$$

其中, \hat{E} 和 \hat{V} 同上。

在优比为常量的无效假设下, T 统计量也近似服从自由度为 $K-1$ 的 χ^2 随机变量渐近分布。

对公共优比的估计: 对于 K 层的四格表, 实际的优比描述为

$$\theta_k = \frac{p_{1k}(1-p_{2k})}{(1-p_{1k})p_{2k}}, \quad k=1, 2, \dots, K$$

并且, 假定实际的公共优比存在, $\theta = \theta_1 = \dots = \theta_K$, 该公共优比的 Mantel 和 Haenszel

的统计量 (1959) 为: $\hat{\theta}_{MH} = \frac{\sum_{k=1}^K \frac{f_{11k}f_{22k}}{n_k}}{\sum_{k=1}^K \frac{f_{12k}f_{21k}}{n_k}}$ 。公共优比的 (自然) 对数为近似正态。

公共优比的渐近置信区间: Robins 等人对 $\log(\hat{\theta}_{MH})$ 给出了一个估计的渐近方差, 它适用于两个渐近情况中

$$\begin{aligned} \hat{\sigma}^2[\log(\hat{\theta}_{MH})] = & \frac{\sum_{k=1}^K \frac{(f_{11k} + f_{22k})f_{11k}f_{22k}}{n_k^2}}{2 \left(\sum_{k=1}^K \frac{f_{11k}f_{22k}}{n_k} \right)^2} \\ & + \frac{\sum_{k=1}^K \frac{(f_{11k} + f_{22k})f_{11k}f_{22k} + (f_{12k} + f_{21k})f_{12k}f_{21k}}{n_k^2}}{2 \left(\sum_{k=1}^K \frac{f_{11k}f_{22k}}{n_k} \right) \left(\sum_{k=1}^K \frac{f_{12k}f_{21k}}{n_k} \right)} \\ & + \frac{\sum_{k=1}^K \frac{(f_{12k} + f_{21k})f_{12k}f_{21k}}{n_k^2}}{2 \left(\sum_{k=1}^K \frac{f_{12k}f_{21k}}{n_k} \right)^2} \end{aligned}$$

$\log(\theta)$ 渐近的 $(100-\alpha)\%$ 的置信区间为

$$\log(\hat{\theta}_{MH}) \pm z(\alpha/2) \hat{\sigma}[\log(\hat{\theta}_{MH})]$$

其中, $z(\alpha/2)$ 是标准正态分布上 $\alpha/2$ 的临界值。

原假设成立的渐近 P 值: 在无效假设 $H_0: \theta_k = \theta_0 (> 0)$ 和双侧的被择假设 $H_1: \theta \neq \theta_0$

下, 使用如下标准正态变量来计算 P 值

$$\Pr\left(|Z| > \left| \frac{\log(\hat{\theta}_{MH}) - \log(\theta_0)}{\hat{\sigma}[\log(\hat{\theta}_{MH})]} \right| \right) = 2 \Pr\left(Z > \left| \frac{\log(\hat{\theta}_{MH}) - \log(\theta_0)}{\hat{\sigma}[\log(\hat{\theta}_{MH})]} \right| \right)$$

作为选择, 我们考虑使用 $\hat{\theta}_{MH}$ 和其估计的精确方差, 它依然同两个限制情况一致:

$$\hat{\theta}^2[\log(\hat{\theta}_{MH})]\hat{\theta}_{MH}^2。$$

则, 渐近 P 值可以用 $\Pr\left(|Z| > \left| \frac{\hat{\theta}_{MH} - \theta_0}{\hat{\sigma}[\log(\hat{\theta}_{MH})]\theta_0} \right| \right)$ 近似。

对这个公式的附加说明, 即使在中等样本量中, $\hat{\theta}_{MH}$ 可能很偏。

(3) 实例

例 7.43 为研究使用别 醇药物会不会引起皮疹, 719 位男性和 605 位女性参加了该试验, 结果见表 7-119, 该表数据已存放在 data07-44.sav 的数据文件中, 试问使用别 醇会不会引起皮疹?

表 7-119 试验结果

性别	皮疹情况	使用别 醇	未使用别 醇
男	引起皮疹	5	36
	未引起皮疹	33	645
女	引起皮疹	10	58
	未引起皮疹	19	518

① 在 SPSS 数据窗口中, 打开数据文件 data07-44.sav。

② 进行加权处理。利用第 2 章 2.3.1.4 中例 2.21 中介绍的方法, 用频数变量做加权处理。

③ 按 Analyze→Descriptive Statistics→Crosstabs, 展开 Crosstabs 对话框, 见图 2-90。在左侧变量名源框中, 选择皮疹情况变量, 单击最上面的右移箭头将其移入 Row(s)框中, 选择使用别 醇变量, 按中间的右移箭头将其移入 Column(s)框中, 选择性别变量, 单击最下面的右移箭头将其移入 Layer 1 of 1 下框中, 把性别设定为第一层(即最里层)的层变量。

表 7-120 样品处理摘要表

	Case Processing Summary					
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
皮疹情况 * 使用别 醇 * 性别	1324	100.0%	0	.0%	1324	100.0%

单击 Statistics 按钮, 打开 Statistics 对话框, 见图 7-12。选择 Cahan's and Mantel-Haenszel Statistics 选项, 在 Test Common odds ratio equals: 后框中保持系统默认值 1, 单击 Continue 按钮返回 Crosstabs 对话框。

④ 单击 OK 按钮运行, 在输出窗口中得到五张输出表格, 见表 7-120、表 7-121、表 7-122、表 7-123 和表 7-124。

⑤ 结果与讨论

表 7-120 为样品处理摘要表, 表 7-121 为按性别分层的交叉描述表, 同表 7-119 中的原始数据。

表 7-122 列出了按性别分层的优比同质性检验结果。两种方法下原假设成立的概率 0.402 大于 0.05, 故不拒绝不同层之间优比值相同的原假设。

表 7-121 按性别分层的交叉描述表

皮疹情况 * 使用别嘌醇情况 * 性别 Crosstabulation

Count			使用别嘌醇情况		Total
性别			使用	未使用	
男	皮疹情况	引起皮疹	5	36	41
		未引起皮疹	33	645	678
	Total		38	681	719
女	皮疹情况	引起皮疹	10	58	68
		未引起皮疹	19	518	537
	Total		29	576	605

表 7-122 分层优比的同质性检验

Tests of Homogeneity of the Odds Ratio			
	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	.703	1	.402
Tarone's	.702	1	.402

表 7-123 分层卡方独立性检验

Tests of Conditional Independence			
	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	19.543	1	.000
Mantel-Haenszel	17.528	1	.000

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

表 7-124 Mantel-Haenszel 的优比估计

Mantel-Haenszel Common Odds Ratio Estimate			
Estimate			3.756
ln(Estimate)			1.323
Std. Error of ln(Estimate)			.318
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	2.016
		Upper Bound	6.998
	ln(Common Odds Ratio)	Lower Bound	.701
		Upper Bound	1.946

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

表 7-123 为分层卡方独立性检验表, 两种方法下皮疹与使用别嘌醇间独立的原假设下, 出现目前统计量的值或者更极端值的概率为 0.000, 表明在去除性别的混杂作用后, 皮疹的出现与使用别嘌醇有关。

表 7-124 为 Mantel-Haenszel 的优比估计, 从上到下依次列出的优比值, 优比值的自

然对数值, 优比值的自然对数值的标准误, 在优比等于 1 的原假设下, 出现目前统计量的值或者更极端值的双侧概率值以及优比值和优比值的自然对数值的 95% 的置信区间。

由于在优比等于 1 的原假设下, 出现目前统计量的值或者更极端值的概率的概率值为 0.000, 因此, 结论同表 7-123 中得到的是一致的。

由于优比值为 3.756, 它表明在去除了性别的混杂效应后, 使用别 的被试对象出现皮疹的风险度大约为未使用别 的被试对象的 3.756 倍。

⑥ 结论

皮疹的出现与使用别 有关, 使用别 的被试对象出现皮疹的风险度大约为未使用别 的被试对象的 3.756 倍。

第 8 章 多因素方差分析和协方差分析

在第 6 章中，我们已经讨论了部分单因变量和多因变量的单因素方差分析，单因素试验资料的统计分析，只能解决一个因素各水平之间的比较问题，但在实际研究中，实验的安排可能更加复杂，所要讨论的影响因变量的因素也可能不止一个，更多的是为两个或多个，例如，在农业试验中，我们知道作物的产量不但同品种有关，还同施肥量、播种期等因素有关，在体育实践中，我们同样也获知，运动成绩的好坏除同运动员本身的条件有关外，还受到训练的运动量、运动强度和运动持续时间等因素的影响，此时在满足一定的前提条件下，用来进行分析的统计方法为多因素方差分析和协方差分析。

同单因素方差分析一样，做多因素方差分析和协方差分析时，理论上也要求需要满足：各观测样本是独立的，各组方差齐性及各组因变量服从正态分布。

采用随机抽样可以保证观测的独立，而用前面介绍的 Levene 统计量可以检验方差的齐性，而对因变量的正态分布的检验可以用探索分析中介绍的数据资料正态性检验的方法，按自变量各个水平分组进行。

由于在多因素方差分析中，正态性和方差齐性的考察是以因素水平组合的单元格为基本单位的，此时，由于实验条件、经费等因素的制约，每个因素水平组合中的观测数不可能很多，所以很难检验，只要数据分布不是明显偏态，不存在极端值，一般都不考虑这两个问题。

方差是否齐性会影响到均值多重比较时检验方法的选择，对方差分析的结果没有太大影响。

对于协方差分析而言，除需满足上述三个前提条件外，还要满足协变量与因变量之间要有线性相关关系，以及每组协变量回归斜率要相等。对于相关关系的检验，可以用散点图或第 10 章相关与回归分析中介绍的方法来加以判定，而对于每组协变量回归斜率要相等的检验，可以通过自变量与协变量的交互作用的 F 检验来判定， F 检验无显著性意义，表明回归斜率齐性。

在 SPSS 的一般线性模型（General Linear Model）过程中，可以用来处理两因素方差分析，多因素方差分析、多元方差分析、重复测量方差分析、协方差分析等。

多因素方差分析和协方差分析涉及的内容很多，本章只对几种较为常用的实验设计下的方差分析和协方差分析方法进行介绍。

8.1 单因变量单因素嵌套设计中的方差分析

8.1.1 单因变量单因素嵌套设计的基本概述

1. 概念

当在单因素完全随机试验所分的若干个组中，每组再分几个亚组（也称子组），每个亚组中有若干个观察值时，称这种试验设计为嵌套设计（nested design）。

在嵌套设计中，各个试验因素对因变量的影响有主次之分，由于次要因素的各个水平嵌套在主要因素的水平之下，因而不能分析它们之间的交互作用。

2. 数据模式

设单因素有 l 组处理（水平），每组处理有 m 个亚组，每个亚组有 n 个观察值，则该试验共有 lmn 个观察值，若用 x_{ijk} 表示观察值，则数据资料有表 8-1 所示的模式。

表 8-1 单因变量单因素嵌套设计中试验资料数据模式

组 l	亚组 m	观察值					
		1	2	...	k	...	n
1	11	x_{111}	x_{112}	...	x_{11k}	...	x_{11n}

	1j	x_{1j1}	x_{1j2}	...	x_{1jk}	...	x_{1jn}

	1m	x_{1m1}	x_{1m2}	...	x_{1mk}	...	x_{1mn}
...
i	$i1$	x_{i11}	x_{i12}	...	x_{i1k}	...	x_{i1n}

	ij	x_{ij1}	x_{ij2}	...	x_{ijk}	...	x_{ijn}

	im	x_{im1}	x_{im2}	...	x_{imk}	...	x_{imn}
...
l	$l1$	x_{l11}	x_{l12}	...	x_{l1k}	...	x_{l1n}

	lm	x_{lm1}	x_{lm2}	...	x_{lmk}	...	x_{lmn}

3. 方差分析方法

设用 T_{ij} 表示每行（亚组）总和，用 T_i 表示每组总和，用 \bar{x}_{ij} 表示每行（亚组）的均值，

用 \bar{x}_i 表示每组均值, 而用 \bar{x} 表示总均值。

在本试验中, 试验资料的每一个观察值的线性可加模型可表示为

$$x_{ijk} = \mu + \tau_i + \varepsilon_{ij} + \delta_{ijk} \quad (i=1,2,\dots,l; j=1,2,\dots,m; k=1,2,\dots,n)$$

式中, μ 为总体均数, τ_i 为组效应 (或处理效应), 可以是固定模型 (指各处理的平均效应 $\tau_i = \mu_i - \mu$, 是一个固定的常量, 同时满足 $\sum \tau_i = 0$ 或 $\sum n_i \tau_i = 0$, 但这个常量未知。本模型的研究对象是处理本身, 处理效应 τ_i 为固定的处理效应。其目的仅在于了解处理间的不同效应), 也可以是随机模型 [指各处理效应 τ_i 不是常量, 而是从正态总体 $N(0, \sigma_\tau^2)$ 中得到的一个随机变量; 总体方差 σ_τ^2 是重要的研究对象, 其目的是要对所研究处理所属的总体做出推论], ε_{ij} 为同组中各亚组的效应, 遵从 $N(0, \sigma_\varepsilon^2)$, δ_{ijk} 为同一亚组中各观察值的随机变异, 服从 $N(0, \sigma^2)$ 。上式说明, 在本试验设计的安排下, 得到的试验资料的每一个观察值的总变异可分解为组间变异、同一组内亚组间变异和同一亚组内各观察值的变异 3 种。

由此, 可将代表变异的总偏差平方和 $SS_T = \sum_{i=1}^l \sum_{j=1}^m \sum_{k=1}^n (x_{ijk} - \bar{x})^2$ 分解为组间平方和 $SS_i = mn \sum_{i=1}^l (\bar{x}_i - \bar{x})^2$ 、组内亚组间平方和 $SS_{m(i)} = \sum_{i=1}^l \sum_{j=1}^m n (\bar{x}_{ij} - \bar{x}_i)^2$ 与误差平方和 $SS_e = \sum_{i=1}^l \sum_{j=1}^m \sum_{k=1}^n n (x_{ijk} - \bar{x}_{ij})^2$ 之和。

在单因变量单因素嵌套试验设计中, 有两种模型, 一个是随机模型, 另一个是混合模型 (称模型中既包含固定模型的试验因素, 又包含随机模型的试验因素的模型为混合模型), 它们的期望均方见表 8-2。

表 8-2 单因变量单因素嵌套试验设计资料的期望均方

变异来源	自由度	期望均方	
		混合模型	随机模型
组间	$l-1$	$\sigma^2 + n\sigma_\varepsilon^2 + mn\kappa_\tau^2$	$\sigma^2 + n\sigma_\varepsilon^2 + mn\sigma_\tau^2$
组内亚组间	$l(m-1)$	$\sigma^2 + n\sigma_\varepsilon^2$	$\sigma^2 + n\sigma_\varepsilon^2$
亚组内	$lm(n-1)$	σ^2	σ^2
总变异	$lmn-1$		

表中, l 表示组数 (处理的水平数), m 表示亚组数, n 表示每个亚组的观测值数, σ^2

表示随机误差的方差, κ_{τ}^2 表示固定效应方差, σ_{τ}^2 随机效应方差, σ_{ε}^2 表示同组中各亚组效应方差。

表 8-2 为检验组间效应和组内亚组间效应提供了检验方案, 在检验各组处理间的不同效应时, 检验假设 $H_0: \kappa_{\tau}^2 = 0$, 或 $\sigma_{\tau}^2 = 0$, 究竟设哪一个, 这取决于试验的模型, 其含义等同于各组处理的均值相等。在做方差分析时, 要用组内亚组间的均方作分母。在检验各亚组间的不同效应时, 其检验的假设 $H_0: \sigma_{\varepsilon}^2 = 0$, 在做方差分析时, 要用亚组内的随机误差的均方作分母。

因此, 可以得到检验用的方差分析表, 见表 8-3。

表 8-3 方差分析表

变异来源	平方和	自由度	均方	F 值
组间	SS_{τ}	$l-1$	$s_{\tau}^2 = \frac{SS_{\tau}}{l-1}$	$F = \frac{s_{\tau}^2}{s_{m(l)}^2}$
组内亚组间	$SS_{m(l)}$	$l(m-1)$	$s_{m(l)}^2 = \frac{SS_{m(l)}}{l(m-1)}$	$F = \frac{s_{m(l)}^2}{s_{\varepsilon}^2}$
亚组内	SS_{ε}	$lm(n-1)$	$s_{\varepsilon}^2 = \frac{SS_{\varepsilon}}{lm(n-1)}$	
总变异	SS_T	$lmn-1$		

注意, 本方差分析法同后面要见到的双因素方差分析是有明显区别的。切勿误用。当检验结果有显著性差异时, 需要进行多重比较。

如果需要做组间均值的多重比较, 则其平均值的标准误为 $s_{\bar{x}} = \sqrt{s_{m(l)}^2 / mn}$ 。而在 LSD 法中, 平均值差异的标准误为 $s_{\bar{x}_1 - \bar{x}_2} = \sqrt{2s_{m(l)}^2 / mn}$ 。

如果需要做组内亚组间均值的多重比较, 则其平均标准误为 $s_{\bar{x}_{ij}} = \sqrt{s_{\varepsilon}^2 / n}$ 。

多重比较的方法有很多种, 在第 6 章 6.2.4.5 中已经做过详细的介绍, 在此不再重述。

8.1.2 单因变量单因素嵌套设计实例分析

例 8.1 为研究油菜种子包衣剂对油菜生长的影响, 用 A、B、C、D 四种油菜种子包衣剂处理油菜品种农杂 62 的种子, 各种子包衣剂处理播种三盒, 采用完全随机设计。播种 20 天后每盒测定 5 株苗高 (cm)。结果如表 8-4 所示, 这些试验结果也已存放在数据文件 data08-01.sav 中, 试比较不同种子对油菜苗高的影响有无显著差异。(数据来源: 中国农业出版社 2007 年 8 月 金益《试验设计与统计分析》第一版 P170)

表 8-4 种子包衣剂对油菜苗高的影响结果

种子包衣剂	种子盒号	苗高测定值 (cm)				
A	A1	6.5	7.3	7.2	6.5	6.0
	A2	7.3	6.0	6.5	6.4	6.1
	A3	6.2	5.7	7.0	6.5	6.9
B	B1	6.3	5.1	6.0	5.8	6.0
	B2	6.2	5.5	5.2	6.0	5.7
	B3	5.1	5.7	6.2	5.5	5.4
C	C1	6.5	8.4	6.5	7.2	6.8
	C2	6.5	6.8	7.0	7.2	7.4
	C3	7.0	6.7	7.5	6.3	6.9
D	D1	5.0	5.2	5.5	5.1	5.6
	D2	5.5	5.0	4.8	4.9	5.7
	D3	4.8	5.4	5.0	4.5	5.8

【题析】本例中的油菜种子包衣剂是研究的主要因素，而每种包衣剂处理随机播种三盒因素是嵌套在各种包衣剂因素中的，是影响油菜生长的次要因素，故本例属于典型的单因变量单因素嵌套设计模型。在 SPSS 中，解类似本题试验设计的题型需要在 General Linear Model 过程的 GLM Univariate 程序中进行，但值得一提的是，它不是双因素分析的原型，因而在 Univariate 对话框中无法直接实现，需要对在 Univariate 对话框中生成的程序，在编辑窗口中作进一步修改后才能得到正解。

单因变量单因素嵌套设计模型数据资料的方差分析方法，在 SPSS 中的具体实现过程如下：

1. 在数据编辑窗口，打开 data08-01.sav。
2. 对 4 种油菜种子包衣剂处理油菜品种农杂 62 的种子组的苗高数据资料进行正态性检验。

各组内数据资料的正态性检验可参照用第 2 章 2.4 节探索分析中介绍的步骤来进行，在第 6 章例 6.19 中也有类似的检验，本例对这一检验过程略，读者可自行验证本例各组内数据资料不拒绝正态分布的原假设。

3. 利用 SPSS 中 GLM Univariate 对话框生成基本程序

(1) 按 Analyze → General Linear Model → Univariate 顺序，打开 Univariate 主对话框，见图 8-1。

在左侧的源变量框中，选中苗高变量，将其移入到 Dependent Variables 框中，作为因变量。

因为本例主要讨论 4 种油菜种子包衣剂对苗高的影响，因此，它属于固定因素（固定因素是指该因素在样本中所有可能的水平都体现了），故选择种子包衣剂变量，将其移入到 Fixed Factor[s]框中。

各种子包衣剂处理播种三盒，采用的是完全随机设计，因此它属于随机因素（随机因素是指该因素在样本中所有可能的水平没有都体现，或不可能都体现。），故选择盒子号变量，将其移入到 Random Factor[s]框中。

因此，本例属于混合模型。

(2) 单击 Model 按钮，打开 Model 对话框，见图 8-2。在 Specify Model 选项中，选择 Custom 选项，要求对模型进行自定义。

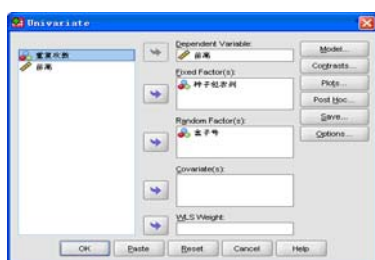


图 8-1 Univariate 主对话框



图 8-2 Model 对话框

在左侧因素和协变量框中，选择种子包衣剂和盒子号变量将其移入到 Model 框中，并在 Build Term[s] 的 Type 中选择 Main effect，要求模型中只包含主效应，其他采用系统默认选项，不作任何选择。单击 Continue 按钮，返回 Univariate 主对话框。

(3) 单击 Paste 按钮，则在语句输出窗口中，生成如图 8-3 所示的程序。

程序清单如下：

UNIANOVA 苗高 BY 种子包衣剂 盒子号

/RANDOM=盒子号

/METHOD=SSTYPE(3)

/INTERCEPT=INCLUDE

/CRITERIA=ALPHA(0.05)

/DESIGN=种子包衣剂 盒子号。

4. 进行方差分析

在语句输出窗口中将其改成如图 8-4 所示的程序，即在最后一行盒子号后面，句号前面加上“(种子包衣剂)”双引号中的内容。注意，括号应是 ASCII 码状态下的括号。它表示“盒子号”是嵌套在“种子包衣剂”中的变量。最后一个语句改为：

/DESIGN=种子包衣剂 盒子号(种子包衣剂)。

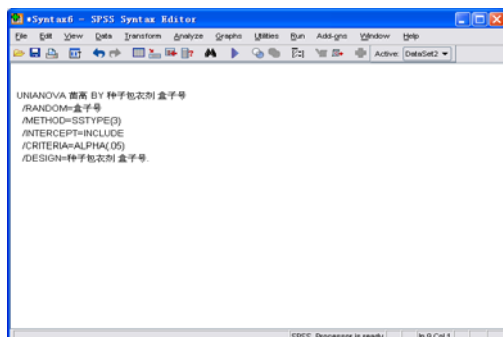


图 8-3 生成的原程序

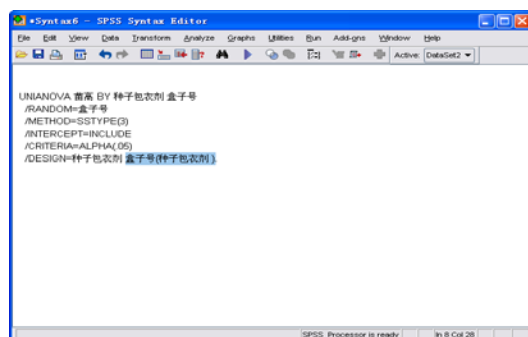


图 8-4 修改最后一行后的程序

单击语句输出窗口中的 Run (见图 8-5) 下拉式菜单中的 All, 则在结果输出窗中, 得到三张表, 其中中间一张为因素之间的方差分析表, 见表 8-5, 它有用。另两张表不用多费心。

5. 结果与讨论

表 8-5 的注解表明, 对种子包衣剂主因素检验时采用的是二级单位误差 (盒子号), 而对次要因素亚组的盒子号检验时采用的是随机误差, 这同表 8-3 中的内容相一致, 正是我们所期望的结果。在表 8-4 中, 表头列出的从左向右依次为变异来源、第三种类型变异分解法得到的变异平方和、自由度、均方、F 值、显著性水平。表中第一行为截距 (假设、误差)、第二行为主因素种子包衣剂 (假设、误差)、第三行为嵌套因素盒子号 (假设、误差)。F 检验表明, 在主因素种子包衣剂均值间无差异的原假设下, 出现目前统计量的值或者更极端值的概率为 0.000, 小于 0.05, 在种子包衣剂内盒间均值无差异的原假设下, 出现目前统计量的值或者更极端值的概率为 0.966, 大于 0.05。

表 8-5 方差分析表

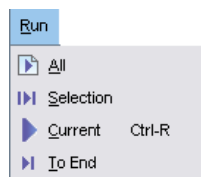


图 8-5 Run 菜单

Tests of Between-Subjects Effects					
Dependent Variable: 苗高					
Source		Type III Sum of Squares	df	Mean Square	Sig.
Intercept	Hypothesis	2236.261	1	2236.261	3.297E4
	Error	.543	8	.068 ^a	.000
种子包衣剂	Hypothesis	29.274	3	9.758	143.852
	Error	.543	8	.068 ^a	.000
盒子号(种子包衣剂)	Hypothesis	.543	8	.068	.291
	Error	11.172	48	.233 ^a	.966

a. MS(盒子号(种子包衣剂))

b. MS(Error)

6. 结论

不同种子包衣剂间苗高有极显著性差异, 而在同一种包衣剂内盒间平均苗高无显著性差异。

7. 各组间均值的多重比较

在 SPSS 中要解决在单因素嵌套设计中获取的主因素各组间数据资料均值的多重比

较是比较困难的,因为它不能在对话框中直接实现,也不能简单地通过修改程序来完成,只能在数据转换菜单计算变量过程中,根据多重比较的基本原理,一步步来实现。

用 LSD 法进行多重比较的具体步骤如下:

(1) 利用多重比较程序来计算各组间的均值之差

① 重复上述 3. (1) 中的步骤,来设置因变量、固定因素和随机因素。

② 单击 Post Hoc 按钮,展开 Post Hoc 多重比较对话框,见图 8-6。

选择等方差假定前提下的 LSD 选项,用 LSD 检验法进行均值间差异的比较检验,以此来获得各组均值之间的差异值。单击 Continue 按钮,返回 Univariate 主对话框。

③ 单击 OK 按钮运行,则在输出窗中,得到 4 张表,其中与此有关的信息在最后一张表中,见表 8-6。

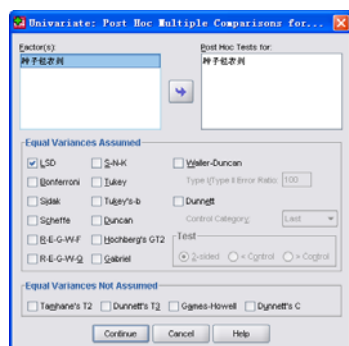


图 8-6 Post Hoc 多重比较对话框

表 8-6 均值之差计算结果

Multiple Comparisons						
前类 后类		Mean Difference (I- J)	Std. Error	Sig.	95% Confidence Interval	
前类 文类	后类 文类				Lower Bound	Upper Bound
A	B	.8267 [*]	.17616	.000	.4725	1.1809
	C	-.4400 [*]	.17616	.016	-.7942	-.0858
	D	1.3533 [*]	.17616	.000	.9991	1.7075
B	A	-.8267 [*]	.17616	.000	-1.1809	-.4725
	C	-1.2667 [*]	.17616	.000	-1.6209	-.9125
	D	.5267 [*]	.17616	.004	.1725	.8809
C	A	.4400 [*]	.17616	.016	.0858	.7942
	B	1.2667 [*]	.17616	.000	.9125	1.6209
	D	1.7933 [*]	.17616	.000	1.4391	2.1475
D	A	-1.3533 [*]	.17616	.000	-1.7075	-.9991
	B	-.5267 [*]	.17616	.004	-.8809	-.1725
	C	-1.7933 [*]	.17616	.000	-2.1475	-1.4391

Based on observed means.
The error term is Mean Square(Error) = .233.

*. The mean difference is significant at the 0.05 level.

Based on observed means.
The error term is Mean Square(Error) = .233.
*. The mean difference is significant at the 0.05 level.

(2) 用 SPSS Transform 菜单的 Compute Variable 过程计算各组均值间的多重比较结果

① 在上面得到表 8-6 的输出窗中,复制均值差值这一列数据,并将其粘贴到新打开的数据编辑窗口中,并作初步的整理,具体形式见图 8-7,这些数据及其后续的计算结果存放在数据文件 data08-02.sav 中。

② 计算平均值差异的标准误

按 Transform→Compute Variable 顺序,展开 Compute Variable 对话框,在 Target Variable 下框中输入目标变量名为“平均值差异的标准误”,在 Numeric Expression 下框中输入计算公式 $s_{\bar{x}_i - \bar{x}_j} = \sqrt{2s_m^2(l)/mn}$ 的表达式为“sqrt(2*0.0678/15)”,单击 OK 按钮,则在数据编辑窗口出现新的一个变量“平均

	种子包衣剂i	种子包衣剂j	均值间差值	var
1	A	B	0.9267	
2	A	C	-0.4400	
3	A	D	1.3533	
4	B	A	-0.9267	
5	B	C	-1.2660	
6	B	D	0.5267	
7	C	A	0.4400	
8	C	B	1.2667	
9	C	D	1.7933	
10	D	A	-1.3530	
11	D	B	-0.5267	
12	D	C	-1.7930	

图 8-7 复制、整理后的数据形式

值差异的标准误”及其计算结果。

③ 计算两均值间无差异成立的概率值

重复②，由于均值间差值与平均值差异标准误之比服从自由度为 $df_{m(l)}$ （亚组自由度）的 t 分布，因此在 Target Variable 下框中输入目标变量名为“概率”，在 Numeric Expression 下框中输入计算公式的表达式为“(1-CDF.T(abs(均值间差值)/平均值差异的标准误,8))*2”，则在数据编辑窗口出现新的一个变量“概率”及其计算结果。

表 8-7 多重比较结果

Case Summaries					
	种子包衣剂	种子包衣剂	均值间差值	平均值差异的标准误	概率
1	A	B	.8267	.0951	.000024
2	A	C	-.4400	.0951	.001693
3	A	D	1.3533	.0951	.000001
4	B	A	-.8267	.0951	.000024
5	B	C	-1.2660	.0951	.000001
6	B	D	.5267	.0951	.000548
7	C	A	.4400	.0951	.001693
8	C	B	1.2667	.0951	.000001
9	C	D	1.7933	.0951	.000000
10	D	A	-1.3530	.0951	.000001
11	D	B	-.5267	.0951	.000548
12	D	C	-1.7930	.0951	.000000
Total	N	12	12	12	12

④ 输出多重比较结果

按 Analyze→Reports→Case Summaries 顺序展开 Case Summaries 对话框，将左侧变量名列表中的全部 5 个变量名移入到 Variables 下框中，单击 OK 按钮运行，则在输出窗口中出现多重比较表，见表 8-7。

⑤ 结论

表 8-7 的多重比较结果表明：在均值间两两比较的结果无差异的原假设下，出现目前统计量的值或者更极端值的概率均小于 0.01，说明 4 种油菜种子包衣剂间有极显著性差异，其中，种子包衣剂 C 的苗高极显著地高于种子包衣剂 A、B 和 D；种子包衣剂 A 的苗高极显著地高于种子包衣剂 B 和 D；种子包衣剂 B 的苗高极显著地高于种子包衣剂 D。

有的学者认为本例的嵌套情形，也可以使用 SPSS 的混合模型更简单地得到解决，但事实上，不但是在 F 检验中，而且在多重比较中都不可能得到同样的真解，读者可自行在 SPSS 的混合模型中进行尝试。

8.2 单因变量单因素随机区组设计中的方差分析

8.2.1 单因变量单因素随机区组设计的基本概述

1. 概念

单因素随机区组设计，最初由 R.A.Fisher 在他的名著《实验设计法》(The Design of Experiments,1935)中引进，它将试验地按土壤肥力梯度划分成等于重复次数的几个区组，区组内的环境条件相对一致，区组内各个小区随机排列。区组间彼此相对差异较大，要求区组方向与土壤肥力梯度垂直，小区的长边与土壤肥力梯度平行。由于实验设计起源于农业试验，所以有区组这种名称。

当每个重复内包含了同一试验的所有处理时称它为一个完全区组。一个区组排成一

排或多排，同一区组的各种处理必须设置在同一直线上。区组不作为影响试验指标的因素，它是被试验严格控制的条件。

2. 数据模式

设单因素随机区组试验中，有 k 个处理， n 个区组，则试验共有 kn 个观察值，若用 x_{ij} 表示观察值，则试验资料的数据结构可用表 8-8 来表示。

表 8-8 单因素随机区组试验中观察资料的数据结构 ($i=1,2,\dots,k; j=1,2,\dots,n$)

品种	区组			
	I	II	...	n
1	x_{11}	x_{12}	...	x_{1n}
2	x_{21}	x_{22}	...	x_{2n}
\vdots	\vdots	\vdots	\vdots	\vdots
i	x_{i1}	x_{i2}	...	x_{in}
\vdots	\vdots	\vdots	\vdots	\vdots
k	x_{k1}	x_{k2}	...	x_{kn}

3. 方差分析方法

设用 $\bar{x}_{i\cdot}$ 表示每行（每个品种）的均值，用 $\bar{x}_{\cdot j}$ 表示每列（每个区组）的均值，而用 \bar{x} 表示总均值。

则在单因素随机区组试验中，试验资料的每一个观察值的线性可加模型可表示为

$$x_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \quad (i=1,2,\dots,k; j=1,2,\dots,n)$$

式中， μ 为总体均数， τ_i 为处理效应（品种效应），可以是固定模型或随机模型，在固定模型中，假定 $\sum \tau_i = 0$ ，而在随机模型中，假定 $\tau_i \sim N(0, \sigma_\tau^2)$ 。 β_j 为区组效应，一般为随机模型，假定 $\beta_j \sim N(0, \sigma_\beta^2)$ ，若它为固定模型，则假定 $\sum \beta_j = 0$ 。 ε_{ij} 为相互独立的随机误差，服从 $N(0, \sigma^2)$ 。

上式说明，在本试验设计的安排下，得到的试验资料的每一个观察值的总变异可分解为处理变异、区组变异和各观察值的变异 3 种。

由此，可将代表变异的总偏差平方和 $SS_T = \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x})^2$ 分解为处理（或品种）间平方和 $SS_t = \sum_{i=1}^k \sum_{j=1}^n (\bar{x}_{i\cdot} - \bar{x})^2$ 、区组间平方和 $SS_r = \sum_{i=1}^k \sum_{j=1}^n (\bar{x}_{\cdot j} - \bar{x})^2$ 与误差平方和 $SS_e = \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2$ 之和。

在单因变量单因素随机区组试验设计中，一般有三种模型，即固定模型、随机模型和混合模型，它们的期望均方见表 8-9。

表 8-9 单因变量单因素随机区组试验设计资料的期望均方

变异来源	自由度	期望均方			
		固定模型	随机模型	混合模型	
		(区组、处理均固定)	(区组、处理均随机)	(区组随机、处理固定)	(区组固定、处理随机)
区组间	$n-1$	$\sigma^2 + k\kappa_\beta^2$	$\sigma^2 + k\sigma_\beta^2$	$\sigma^2 + k\sigma_\beta^2$	$\sigma^2 + k\kappa_\beta^2$
处理间	$k-1$	$\sigma^2 + n\kappa_\tau^2$	$\sigma^2 + n\sigma_\tau^2$	$\sigma^2 + n\kappa_\tau^2$	$\sigma^2 + n\sigma_\tau^2$
试验误差	$(k-1)(n-1)$	σ^2	σ^2	σ^2	σ^2

表中， n 表示区组数， k 表示处理的水平数， σ^2 表示随机误差的方差， κ_τ^2 表示固定模型中处理效应方差， σ_τ^2 表示随机模型中处理效应方差， κ_β^2 表示固定模型中区组效应方差， σ_β^2 表示随机模型中区组效应方差。

表 8-9 为检验区组间效应和处理间效应提供了检验方案，在检验各组处理间的不同效应时，检验假设 $H_0: \kappa_\tau^2 = 0$ ，或 $\sigma_\tau^2 = 0$ ，究竟设哪一个，这取决于试验的模型，其含义等同于各组处理的均值相等。在做方差分析时，要用试验误差的均方作分母。在检验各区组间的不同效应时，其检验的假设 $H_0: \sigma_\beta^2 = 0$ ，或 $H_0: \kappa_\beta^2 = 0$ ，在做方差分析时，也要用试验误差的均方作分母。

因此，可以得到检验用的方差分析表，见表 8-10。

表 8-10 方差分析表

变异来源	平方和	自由度	均方	F 值
区组间	SS_r	$n-1$	$s_r^2 = \frac{SS_r}{n-1}$	$F = \frac{s_r^2}{s_e^2}$
处理间	SS_t	$k-1$	$s_t^2 = \frac{SS_t}{k-1}$	$F = \frac{s_t^2}{s_e^2}$
误差	SS_e	$(k-1)(n-1)$	$s_e^2 = \frac{SS_e}{(k-1)(n-1)}$	
总变异	SS_T	$kn-1$		

当检验结果有显著性差异时，需要进行多重比较。方法同第 6 章单因素方差分析部分所介绍的方法是一致的。

8.2.2 单因变量单因素随机区组设计实例分析

例 8.2 为了比较正确地评介分别含 5 种不同蛋白质饲料的营养价值，研究人员随机

选用 8 窝大白鼠，每窝选取性别相同、体重接近的 5 只，用随机的方法将每窝中的 5 只分到 5 个饲料组中，饲养 9 周后，测得它们的体重增加量，结果见 data08-03.sav。试问 5 种饲料的营养价值之间有无显著性差异？

在 SPSS 中的具体实现过程如下：

(1) 在数据编辑窗口，打开 data08-03.sav。

(2) 按 Analyze → General Linear Model → Univariate 顺序，打开 Univariate 主对话框，见图 8-1。

在左侧的变量源框中，选中 **体重增加量** 变量，将其移入到 Dependent Variables 框中，作为因变量。

因为本例主要讨论 5 种不同蛋白质饲料对体重的影响，因此，品种饲料属于固定因素，故选择 **饲料** 变量，将其移入到 Fixed Factor[s] 框中。

由于 8 窝大白鼠是随机选取的，故选择 **窝别** 变量，将其移入到 Random Factor[s] 框中。

因为本例中既包含有固定模型的试验因素，又包含有随机模型的试验因素，因此，它属于混合模型。

(3) 单击 Model 按钮，打开 Model 对话框，见图 8-2。在 Specify Model 选项中，选择 Custom 选项，要求对模型进行自定义。

在左侧因素和协变量框中，选择 **饲料** 和 **窝别** 变量将其移入到 Model 框中，并在 Build Term[s] 的 Type 中选择 Main effect，要求模型中只包含主效应，其他采用系统默认选项，不作任何选择。单击 Continue 按钮，返回 Univariate 主对话框。

(4) 单击 OK 按钮运行，则在输出窗中，得到三张表，其中中间一张为因素之间的方差分析表，见表 8-11，另两张表不用多费心。

(5) 结果与讨论

表 8-11 的表头内容同表 8-5 一样，从表 8-11 中可见，在饲料间无差异的原假设下，出现目前统计量的值或者更极端值的概率为 0.000，小于 0.05，故拒绝饲料间无差异的原假设，而认为至少有两个饲料间对体重的影响是不一样的；而在窝别间无差异的原假设下，出现目前统计量的值或者更极端值的概率为 0.122，大于 0.05，故不拒绝这个原假设，即认为窝别间无显著性差异。

(6) 结论

不同饲料间的大白鼠体重增加量有极显著性差异，但不同窝别之间无显著性差异。

表 8-11 方差分析表

Tests of Between-Subjects Effects						
Dependent Variable: 体重增加量						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	95257.600	1	95257.600	998.806	.000
	Error	667.600	7	95.371 ^a		
饲料	Hypothesis	12004.150	4	3001.038	57.450	.000
	Error	1462.650	28	52.238 ^b		
窝别	Hypothesis	667.600	7	95.371	1.826	.122
	Error	1462.650	28	52.238 ^b		

a. MS(窝别)

b. MS(Error)

(7) 不同饲料间均值的多重比较

单击 Post Hoc 按钮, 展开 Post Hoc 多重比较对话框, 见图 8-6。在 Factor(s) 下框中, 将饲料变量移入到 Post Hoc Tests for 下框中, 由于本例的饲料品种不太多, 故本例选择等方差假定前提下的 LSD 选项, 用 LSD 检验法进行多重比较。单击 Continue 按钮, 返回 Univariate 主对话框。

表 8-12 不同饲料体重增加量均值间的多重比较

因变量	饲料	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
甲	乙	17.0000 ^a	3.61378	.000	9.5975	24.4025
	丙	7.3750 ^a	3.61378	.051	-.0275	14.7775
	丁	-12.1250 ^b	3.61378	.002	-19.5275	-4.7225
	戊	-33.1250 ^b	3.61378	.000	-40.5275	-25.7225
乙	甲	-17.0000 ^b	3.61378	.000	-24.4025	-9.5975
	丙	-9.6250 ^a	3.61378	.013	-17.0275	-2.2225
	丁	-29.1250 ^b	3.61378	.000	-36.5275	-21.7225
	戊	-50.1250 ^b	3.61378	.000	-57.5275	-42.7225
丙	甲	-7.3750 ^a	3.61378	.051	-14.7775	.0275
	乙	9.6250 ^a	3.61378	.013	2.2225	17.0275
	丁	-19.5000 ^b	3.61378	.000	-26.9025	-12.0975
	戊	-40.5000 ^b	3.61378	.000	-47.9025	-33.0975
丁	甲	12.1250 ^a	3.61378	.002	4.7225	19.5275
	乙	29.1250 ^a	3.61378	.000	21.7225	36.5275
	丙	19.5000 ^a	3.61378	.000	12.0975	26.9025
	戊	-21.0000 ^b	3.61378	.000	-28.4025	-13.5975
戊	甲	33.1250 ^a	3.61378	.000	25.7225	40.5275
	乙	50.1250 ^a	3.61378	.000	42.7225	57.5275
	丙	40.5000 ^a	3.61378	.000	33.0975	47.9025
	丁	21.0000 ^a	3.61378	.000	13.5975	28.4025

Based on observed means.
The error term is Mean Square(Error) = 52.238.

单击 OK 按钮运行, 则在输出窗中, 得到四张表, 前三张表同原来的三张表一样, 第四张表为不同饲料体重增加量均值间的多重比较表, 见表 8-12。

从表 8-12 可见, 4 种饲料之间, 除在甲与丙之间无差异原假设下, 出现目前统计量的值或者更极端值的概率为 0.051 大于 0.05 以外, 其他在两两间无差异原假设下, 出现目前统计量的值或者更极端值的概率均小于 0.05。戊饲料对体重的增加效果极显著地高于甲、乙、丙和丁饲料; 丁饲料对体重的增加效果极显著地高于甲、乙和丙饲料; 甲饲料对体重的增加效果显著地高于乙和丙饲料; 丙饲料对

体重的增加效果显著地高于乙饲料。

由上可知, 各饲料对体重增加量的影响效果依次为戊、丁、甲、丙和乙饲料。

例 8.3 在一次水稻品种比较研究中, 采用 8 个水稻品种, 其中第 2 种为对照品种, 用随机区组设计, 重复三次, 小区计产面积 40 平方米, 水稻产量的试验结果已存放在数据文件 data08-04.sav 中。试分析各品种的产量间是否有显著差异。

在 SPSS 中的具体实现过程如下:

(1) 在数据编辑窗口, 打开 data08-04.sav。

(2) 按 Analyze → General Linear Model → Univariate 顺序, 打开 Univariate 主对话框, 见图 8-1。

在左侧的变量源框中, 选中产量变量, 将其移入到 Dependent Variables 框中, 作为因变量。

因为本例主要讨论 8 个水稻品种对水稻产量的影响, 因此, 品种属于固定因素, 故选择品种变量, 将其移入到 Fixed Factor[s] 框中。

由于采用的是完全随机区组设计, 故选择区组变量, 将其移入到 Random Factor[s] 框中。

因为本例中既包含有固定模型的试验因素, 又包含有随机模型的试验因素, 因此, 本例属于混合模型。

(3) 单击 Model 按钮, 打开 Model 对话框, 见图 8-2。在 Specify Model 选项中, 选

择 Custom 选项，要求对模型进行自定义。

在左侧因素和协变量框中，选择品种和区组变量将其移入到 Model 框中，并在 Build Term[s] 的 Type 中选择 Main effect，要求模型中只包含主效应，其他采用系统默认选项，不作任何选择。单击 Continue 按钮，返回 Univariate 主对话框。

(4) 单击 OK 按钮运行，则在输出窗中，得到三张表，其中中间一张为因素之间的方差分析表，见表 8-13，另两张表不用多费心。

(5) 结果与讨论

从表 8-13 中可见，在品种间无差异的原假设下，出现目前统计量的值或者更极端值的概率为 0.002，小于 0.05，故拒绝品种间无差异的原假设，而认为至少有两个品种间产量是不一样的；而在区组间无差异的原假设下，出现目前统计量的值或者更极端值的概率为 0.143，大于 0.05，故不拒绝这个原假设，即认为土壤肥力间无显著性差异。

(6) 结论

不同品种间的水稻产量有极显著性差异，但土壤肥力间无显著性差异。

(7) 不同品种间均值的多重比较

单击 Post Hoc 按钮，展开 Post Hoc 多重比较对话框，见图 8-6。在 Factor(s) 下框中，将品种变量移入到 Post Hoc Tests for 下框中，由于本例的品种较多，有 8 种，为使输出表格不要太大，本例选择等方差假定前提下的 S-N-K 选项，用 S-N-K 检验法进行同质性检验。单击 Continue 按钮，返回 Univariate 主对话框。

表 8-14 品种间水稻产量的多重比较

Student-Newman-Keuls			
品种	N	Subset	
		1	2
A7	3	20.4667	
A6	3	20.7000	
A4	3	20.7000	
A8	3	21.2000	
A1	3	22.2000	
A2	3	22.5000	
A3	3	23.2667	
A5	3		26.5000
Sig.		.250	1.000

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 1.966.

7 种品种在的第一组，这说明 A5 品种的水稻产量显著地高于对照组 A2 和其他各组，其他各组和对照组 A2 的均值间无显著性差异。

由于区组间无显著性差异，故区组间不需要做多重比较。

表 8-13 方差分析表

Tests of Between-Subjects Effects					
Dependent Variable: 产量					
Source		Type III Sum of Squares	df	Mean Square	Sig.
Intercept	Hypothesis	11819.282	1	11819.282	2.679E3
	Error	8.823	2	4.412 ^a	.000
品种	Hypothesis	84.665	7	12.095	6.151
	Error	27.530	14	1.966 ^a	.002
区组	Hypothesis	8.823	2	4.412	2.243
	Error	27.530	14	1.966 ^a	.143

a. MS(区组)
b. MS(Error)

8.3 单因变量多因素试验的方差分析

8.3.1 单因变量双因素完全随机试验的方差分析

8.3.1.1 单因变量双因素无重复完全随机试验的方差分析

1. 概念

在单因素随机区组试验中，考察的试验因素只有一个，而区组因素是一个重要的非试验因素，当影响试验结果的试验因素有 2 个时，称它为双因素试验设计。

在双因素试验中，试验因素间关系是平等的，可以认为试验结果的变化是由于两个因素共同作用的结果。它一般地要求 A 因素的每个水平和 B 因素的每个水平均衡相遇，构成交叉分组，当每组只有一个观察值时，称这样的资料为无重复的双因素完全随机试验资料。

需要注意的是，在无重复试验时获取的资料是不能讨论两个因素之间的交互作用的。另外，从方差分析的计算方法上而言，双因素完全随机试验资料的分析方法同单因素随机区组试验资料的分析方法是完全一样的，正因为如此，在许多统计书籍中，将这两个试验设计当作不同名称的同一种试验设计方法看待。

2. 数据模式

在双因素无重复完全随机试验中，设 A 因素有 a 个水平 (A_1, A_2, \dots, A_a)，B 因素有 b 个水平 (B_1, B_2, \dots, B_b)，则试验全部处理共有 ab 个水平组合，每个组合只有一个观察值，若用 x_{ij} 表示观察值，共有 ab 个观察值，其数据结构如表 8-15 所示。

表 8-15 双因素无重复完全随机试验中观察资料的数据结构 ($i=1,2,\dots,a; j=1,2,\dots,b$)

A 因素	B 因素			
	B_1	B_2	...	B_b
A_1	x_{11}	x_{12}	...	x_{1b}
A_2	x_{21}	x_{22}	...	x_{2b}
\vdots	\vdots	\vdots	\vdots	\vdots
A_i	x_{i1}	x_{i2}	...	x_{ib}
\vdots	\vdots	\vdots	\vdots	\vdots
A_a	x_{a1}	x_{a2}	...	x_{ab}

3. 方差分析方法

设用 $\bar{x}_{i\cdot}$ 表示 A 因素每个水平（每行）的均值，用 $\bar{x}_{\cdot j}$ 表示 B 因素每个水平（每列）的均值，而用 \bar{x} 表示总均值。

则在双因素无重复完全随机试验中，试验资料的每一个观察值的线性可加模型可表

示为

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (i=1,2,\dots,a; j=1,2,\dots,b)$$

式中, μ 为总体均数, α_i 、 β_j 分别为 A_i 、 B_j 的效应, 可以是固定模型或随机模型, 在固定模型中, 假定 $\sum \alpha_i = 0$, $\sum \beta_j = 0$, 而在随机模型中, 假定 $\alpha_i \sim N(0, \sigma_A^2)$ 、 $\beta_j \sim N(0, \sigma_B^2)$ 。 ε_{ij} 为相互独立的随机误差, 服从 $N(0, \sigma^2)$ 。

上式说明, 在本试验设计的安排下, 得到的试验资料的每一个观察值同时受到 A 因素、B 因素和随机误差的共同影响。

由此, 可将代表变异的总偏差平方和 $SS_T = \sum_{i=1}^a \sum_{j=1}^b (x_{ij} - \bar{x})^2$ 分解为 A 因素各水平间变异的平方和 $SS_A = \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{i\cdot} - \bar{x})^2$ 、B 因素各水平间变异的平方和 $SS_B = \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{\cdot j} - \bar{x})^2$ 与随机误差平方和 $SS_e = \sum_{i=1}^a \sum_{j=1}^b (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2$ 之和。

在双因变量无重复完全随机试验设计中, 一般有三种模型, 即固定模型、随机模型和混合模型, 它们的期望均方见表 8-16。

表 8-16 单因变量双因素无重复完全随机试验设计资料的期望均方

变异来源	自由度	期望均方			
		固定模型	随机模型	混合模型	
		(A、B 均固定)	(A、B 均随机)	(A 固定、B 随机)	(A 随机、B 固定)
A 因素	$a-1$	$\sigma_e^2 + b\kappa_A^2$	$\sigma_e^2 + b\sigma_A^2$	$\sigma_e^2 + b\kappa_A^2$	$\sigma_e^2 + b\sigma_A^2$
B 因素	$b-1$	$\sigma_e^2 + a\kappa_B^2$	$\sigma_e^2 + a\sigma_B^2$	$\sigma_e^2 + a\sigma_B^2$	$\sigma_e^2 + a\kappa_B^2$
试验误差	$(a-1)(b-1)$	σ_e^2	σ_e^2	σ_e^2	σ_e^2
总变异	$ab-1$				

表中, a 表示 A 因素的水平数, b 表示 B 因素的水平数, σ_e^2 表示试验误差的方差, κ_A^2 表示固定模型中 A 因素效应方差, σ_A^2 表示随机模型中 A 因素效应方差, κ_B^2 表示固定模型中 B 因素效应方差, σ_B^2 表示随机模型中 B 因素效应方差。

从表 8-16 中可见, 在各种不同模型下, 各项变异的期望方差组成分量是相同的, 因此对 A 因素、B 因素效应做 F 检验时, 所作的原假设 $H_0: \kappa_A^2 = 0$ (或 $\sigma_A^2 = 0$)、 $\kappa_B^2 = 0$ (或 $\sigma_B^2 = 0$) , 所用的分母都是相同的, 均以误差项均方 (s_e^2) 做分母。

因此, 可以得到检验用的方差分析表, 见表 8-17。

表 8-17 单因变量双因素无重复完全随机试验设计资料的方差分析表

变异来源	平方和	自由度	均方	F 值
A 因素	SS_A	$a-1$	$s_A^2 = \frac{SS_A}{a-1}$	$F = \frac{s_A^2}{s_e^2}$
B 因素	SS_B	$b-1$	$s_B^2 = \frac{SS_B}{b-1}$	$F = \frac{s_B^2}{s_e^2}$
随机误差	SS_e	$(a-1)(b-1)$	$s_e^2 = \frac{SS_e}{(a-1)(b-1)}$	
总变异	SS_T	$ab-1$		

当检验结果有显著性差异时, 需要进行多重比较。方法同第 6 章单因素方差分析部分所介绍的方法是一致的。

4. 对总偏差平方和的贡献率

由于总试验次数为 $a \times b$ 次, 理应有 $a \times b$ 个实验误差, 否则不匹配。其实, 在 SS_E 中含着 $(a-1)(b-1)$ 个实验误差, 其余 $ab - (a-1)(b-1) = (a-1) + (b-1) + 1$ 个实验误差中有 $(a-1)$ 个实验误差包含于 SS_A 中, 有 $(b-1)$ 个实验误差包含于 SS_B 中, 还有一个实验误差

包括于 CT (修正项) 中, $CT = \frac{(x_{11} + x_{12} + \cdots + x_{ab})^2}{a \times b}$ 。

因此, 可以通过以下的方式来计算因素的纯效应和贡献率:

A 因素的纯效应: $SS'_A = SS_A - (a-1)s_e^2$

B 因素的纯效应: $SS'_B = SS_B - (b-1)s_e^2$

误差 E 的纯效应: $SS'_e = SS_e + (a-1)s_e^2 + (b-1)s_e^2$

贡献率 $\rho(\%)$ 为: $100\% = \rho_A + \rho_B + \rho_E = SS'_A / SS_T + SS'_B / SS_T + SS'_e / SS_T$

5. 多重比较

若用 LSD 法进行多重比较, 则当因素 A 各水平之间进行多重比较时, $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{b}}$,

B 因素各水平间进行多重比较时, $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{a}}$, 两因素水平组合之间进行多重比较时,

$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{n}}$, t 分布的自由度为 $df = (a-1)(b-1)$ 。

6. 实例分析

例 8.4 为了考查 200 米跑训练中, 不同间歇时间的重复跑对运动员体内血乳酸含量影响的情况, 对随机抽取的某队 6 名运动员用 4 中不同间歇时间的重复跑进行试验, 观

察其体内血乳酸含量的变化情况, 实验数据见表 8-18, 数据文件见 data08-05.sav。试分析不同时间的间歇跑训练和运动员个体之间的差异是否影响血乳酸含量的变化。

表 8-18 不同间歇时间的重复跑时运动员体内血乳酸含量

受试者 A_i	重复间歇跑间歇时间 $B_j(\text{min})$			
	$B_1(4')$	$B_2(3')$	$B_3(2')$	$B_4(4', 3', 2')$
1	159	168	177	196
2	161	171	180	199
3	173	185	194	215
4	252	267	281	311
5	232	248	261	289
6	184	183	185	186

【题析】 本例是一个二因素完全随机因素不重复测定的试验资料。A 因素运动员有 6 个水平, B 因素重复间歇跑间歇时间有 4 个水平, 共有 24 个观察值。

在 SPSS 中进行统计分析的具体步骤如下:

(1) 在数据编辑窗口, 打开 data08-05.sav。

(2) 按 Analyze → General Linear Model → Univariate 顺序, 打开 Univariate 主对话框, 见图 8-1。

(3) 在左侧因素和协变量框中, 选择血乳酸变量, 单击向右箭头, 将其送入 Dependent Variable 框; 用同样方法分别将间歇时间、运动员二个因素变量分别送入 Fixed Factor(s) 框和 Random Factor[s] 框中。

(4) 单击 Model 按钮, 打开 Model 对话框, 见图 8-2。在 Specify Model 选项中, 选择 Custom 选项, 要求对模型进行自定义。

在左侧因素和协变量框中, 选择运动员和血乳酸变量将其移入到 Model 框中, 并在 Build Term[s] 的 Type 中选择 Main effect, 要求模型中只包含主效应, 关闭 include intercept in model 选项。单击 Continue 按钮, 返回 Univariate 主对话框。

表 8-19 方差分析表

Tests of Between-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
运动员	Hypothesis	40366.708	5	8073.342	107.995	.000
	Error	1121.458	15	74.764 ^a		
间歇时间	Hypothesis	4998.792	3	1666.264	22.287	.000
	Error	1121.458	15	74.764 ^a		

a. MS(Error)

(5) 单击 OK 按钮运行, 则在输出窗中, 得到三张表, 第二张表为所需的两因素之间的方差分析表, 见表 8-19。

(6) 结果与讨论

从表 8-19 中可见, 在运动员个体间血乳酸值无差异的原假设下, 出现目前统计量的

值或者更极端值的概率为 0.000, 同样在 4 种不同间歇时间间血乳酸值无差异的原假设下, 出现目前统计量的值或者更极端值的概率也为 0.000, 因此, 运动员个体差异因素和不同间歇时间因素都对血乳酸值有极显著性影响。

(7) 不同间歇时间均值的多重比较

单击 Post Hoc 按钮, 展开 Post Hoc 多重比较对话框, 见图 8-6。在 Factor(s) 下框中, 将间歇时间变量移入到 Post Hoc Tests for 下框中, 由于本例只有 4 种不同间歇跑, 故本例选择等方差假定前提下的 LSD 选项, 用 LSD 检验法进行多重比较。按 Continue 返回 GLM Univariate 对话框。

单击 OK 按钮运行, 则在输出窗中, 得到四张表, 前三张表同原来的三张表一样, 第四张表为 4 种不同间歇重复跑血乳酸均值间的多重比较表, 见表 8-20。

表 8-20 LSD 多重比较结果

血乳酸 LSD		Mean Difference (I- J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
4分钟 2分钟 混合时间	3分钟	-10.1667	4.99212	.060	-20.8071	4.738
	2分钟	-18.5000 [*]	4.99212	.001	-30.1405	-8.8595
	混合时间	-38.1667 [*]	4.99212	.000	-49.8071	-28.5262
3分钟 2分钟 混合时间	4分钟	10.1667	4.99212	.060	-4.738	20.8071
	2分钟	-9.3333	4.99212	.081	-19.9738	1.3071
	混合时间	-29.0000 [*]	4.99212	.000	-39.6405	-18.3595
2分钟 混合时间	4分钟	19.5000 [*]	4.99212	.001	8.8595	30.1405
	3分钟	9.3333	4.99212	.081	-1.3071	19.9738
	混合时间	-19.6667 [*]	4.99212	.001	-30.3071	-9.0262
混合时间 4分钟 3分钟 2分钟	4分钟	38.1667 [*]	4.99212	.000	28.5262	49.8071
	3分钟	28.0000 [*]	4.99212	.000	18.3595	39.6405
	2分钟	18.6667 [*]	4.99212	.001	9.0262	30.3071

Based on observed means.
The error term is Mean Square(Error) = 74.764.
*. The mean difference is significant at the 0.05 level.

从表 8-20 可见, 4 种不同间歇重复跑之间, 除间歇 4 分钟与间歇 3 分钟之间和间歇 3 分钟与间歇 2 分钟之间无差异的概率为 0.060 和 0.081 大于 0.05 以外, 其他两两间无差异的概率均小于 0.05。

混合时间重复跑对血乳酸的影响极显著地高于间歇时间为 2 分钟、3 分钟和 4 分钟的重复跑; 间隔 2 分钟的重复跑对血乳酸的影响显著地高于间歇时间 4 分钟的重复跑。

(8) 影响血乳酸值变化的主要因素分析

根据纯效应的计算公式可得

A 因素的纯效应

$$SS'_A = SS_A - (a-1)s_e^2 = 40366.71 - 5 \times 74.76 = 39992.91$$

B 因素的纯效应

$$SS'_B = SS_B - (b-1)s_e^2 = 4998.79 - 3 \times 74.76 = 4774.51$$

误差 E 的纯效应

$$SS'_e = SS_e + (a-1)s_e^2 + (b-1)s_e^2 = 1121.46 + 5 \times 74.76 + 3 \times 74.76 = 1719.54$$

因此, 贡献率 $\rho(\%)$ 为

$$\begin{aligned} 100\% &= \rho_A + \rho_B + \rho_E = SS'_A / SS_T + SS'_B / SS_T + SS'_E / SS_T \\ &= \frac{39992.91}{46486.96} + \frac{4774.51}{46486.96} + \frac{1719.54}{46486.96} \\ &= 86.03\% + 10.27\% + 3.7\% \end{aligned}$$

可以认为,不同受试者、不同间歇时间对血乳酸含量的影响的差异间有非常显著的意义。在对总偏差平方和的贡献中,因素 A 最大为 86.03%,因素 B 次之,为 10.27%,残差的影响最小,为 3.7%。由此可见,血乳酸值的变化主要是由于个体差异的影响所造成的,当然也不能忽视不同间歇时间的重复跑因素的影响。在不同间歇时间因素上,混合时间重复跑对血乳酸的影响最大。

例 8.5 表 8-21 是 1987 年 8 月在成都体院举办的全国体院系统武术邀请赛上,女子通臂拳比赛的临场统计,其中的数据已建成数据文件 data08-06.sav,试用方差分析法对评分性项目的裁判员的裁判水平进行评估。

表 8-21 女子通臂拳比赛临场得分统计

运动员	裁判员			
	1	2	3	4
1	88.6	8.9	8.8	8.8
2	8.9	8.9	8.9	8.7
3	8.8	8.7	8.9	9
4	8.4	8.6	8.6	8.6
5	8.6	8.6	8.6	8.6
6	8.4	8.8	8.8	8.8
7	8.1	8.3	8.5	8.5

【题析】影响打分的因素一是运动员本身的水平因素,另一个是裁判员的水平因素,而试验测试指标为运动员得分,因而本题是一个单因变量双因素的固定模型的方差分析问题。

在 SPSS 中进行统计分析的具体步骤如下:

(1) 在数据编辑窗口,打开 data08-06.sav。

(2) 按 Analyze → General Linear Model → Univariate 顺序,打开 Univariate 主对话框,见图 8-1。

(3) 在左侧因素和协变量框中,选择得分变量,单击向右箭头,将其送入 Dependent Variable 框;用同样方法分别将裁判员、运动员二个因素变量送入 Fixed Factor(s) 框中。

(4) 单击 Model 按钮,打开 Model 对话框,见图 8-2。在 Specify Model 选项中,选择 Custom 选项,要求对模型进行自定义。

在左侧因素和协变量框中,选择裁判员和运动员变量将其移入到 Model 框中,并在 Build Term[s] 的 Type 中选择 Main effect,要求模型中只包含主效应,关闭 include intercept in model 选项。单击 Continue 按钮,返回 Univariate 主对话框。

(5) 单击 OK 按钮运行,则在输出窗中,得到二张表,其中第二张表为所需的两因

素之间的方差分析表，见表 8-22。

表 8-22 方差分析表

Tests of Between-Subjects Effects					
Dependent Variable: 得分					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	2104.635 ^a	10	210.464	1.612E4	.000
裁判员	.153	3	.051	3.894	.026
运动员	.794	6	.132	10.131	.000
Error	.235	18	.013		
Total	2104.870	28			

a. R Squared = 1.000 (Adjusted R Squared = 1.000)

运动员之间的水平是不同的。因而，裁判水平因素和运动员水平因素都对最后得分有显著性影响。

(7) 不同裁判员间得分均值的多重比较

单击 Post Hoc 按钮，展开 Post Hoc 多重比较对话框，见图 8-6。在 Factor(s) 下框中，将裁判员变量移入到 Post Hoc Tests for 下框中，选择等方差假定前提下的 LSD 选项，用 LSD 检验法进行多重比较。单击 Continue 按钮，返回 Univariate 主对话框。

单击 OK 按钮运行，则在输出窗中，得到三张表，第三张表为不同裁判员间得分均值的多重比较表，见表 8-23。

从表 8-23 可见，4 个裁判员之间，1 号裁判员得分最低同其他三名裁判员的得分间有显著性差异，两两间无差异的概率均小于 0.05。2 号、3 号和 4 号裁判员的得分间无显著性差异。

(8) 对打分结果的主要影响因素分析

仿例 8.4 中的做法，可得到表 8-24 所示的各因素和误差的纯效应和贡献率。

表 8-24 各因素的贡献率计算结果表

误差来源	偏差平方和	纯效应	贡献率
运动员	0.7935714	0.715238	60.56%
裁判员	0.1525	0.1133	9.60%
随机误差	0.235	0.3525	29.84%
总	1.18107143		

(6) 结果与讨论

从表 8-22 中可见，在裁判员之间得分无差异的原假设下，出现目前统计量的值或者更极端值的概率为 0.026，小于 0.05，因而可以认为至少有两个裁判之间打分是有差异的；同样在运动员间得分相同的原假设下，出现目前统计量的值或者更极端值的概率为 0.000，说明运动员

表 8-23 不同裁判员得分均值间的多重比较

Multiple Comparisons						
得分 LSD						
① 裁判员	② 裁判员	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-.14 [*]	.061	.031	-.27	-.01
	3	-.19 [*]	.061	.007	-.31	-.06
	4	-.17 [*]	.061	.012	-.30	-.04
2	1	.14 [*]	.061	.031	.01	.27
	3	-.04	.061	.492	-.17	.09
	4	-.03	.061	.646	-.16	.10
3	1	.19 [*]	.061	.007	.06	.31
	2	.04	.061	.492	-.09	.17
	4	.01	.061	.818	-.11	.14
4	1	.17 [*]	.061	.012	.04	.30
	2	.03	.061	.646	-.10	.16
	3	-.01	.061	.818	-.14	.11

Based on observed means.

The error term is Mean Square(Error) = .013.

*. The mean difference is significant at the 0.05 level.

从各因素对总偏差平方和的贡献来看，运动员间的差异是第一位的占总偏差平方和的 60.56%，说明运动员本身的水平是影响打分的最主要的因素，第二位是残差，占总偏差平方和的 29.84%，而裁判水平的差异仅占总偏差平方和的 9.60%，说明裁判的水平是

相当接近的。

(9) 裁判员水平分析

由于裁判员的裁判水平间有显著差异, 因此, 可以讨论如何来判别他们水平高低的问题。

显然, 我们不能用第 j 个裁判员评分的平均值 $\bar{x}_{\cdot j}$ 的大小去评估裁判水平的高低, 因此, 需要引进一个表示裁判水平高低的量。

设第 i 号运动员得分的准确值为 \bar{x}_i^* , 那么, $\bar{x}^* = (\bar{x}_1^*, \bar{x}_2^*, \dots, \bar{x}_a^*)$ 是这组运动员得分的准确值, 第 j 号裁判员对这组运动员的评分 $x_j = (x_{1j}, x_{2j}, \dots, x_{aj})$ 是 \bar{x}^* 的一组观察值。

量 $|x_j - \bar{x}^*| = \sqrt{\sum_{i=1}^a (x_{ij} - \bar{x}_i^*)^2}$ 表示第 j 号裁判员评分的误差。

$|x_j - \bar{x}^*|$ 越小, 裁判员水平越高, 这里由于 \bar{x}^* 不知道, 因此, 必须找出它的一组无偏估计量。

我们记第 i 号运动员的平均得分为: $\bar{x}_i = \frac{1}{b} \sum_{j=1}^b x_{ij}, i=1, 2, \dots, 7$, 根据比赛规则, 第 i 号运动员的实际得分为: $\bar{x}_i' = \frac{1}{b-2} \left[\sum_{j=1}^b x_{ij} - \max(x_{ij}) - \min(x_{ij}) \right], 1 \leq j \leq b$, 运动员的名次就是按 \bar{x}_i' 的大小来定的, 故有表 8-25。

表 8-25 运动员的期望得分与最终得分比较表

运动员编号	1	2	3	4	5	6	7
\bar{x}_i	8.80	8.90	8.85	8.60	8.60	8.80	8.40
比赛名次	3	1	2	6	5	4	7
\bar{x}_i'	8.775	8.85	8.85	8.575	8.60	8.70	8.35
\bar{x}_i 大小	3	1	1	6	5	4	7

由上表可见, 运动员得分的平均值 \bar{x}_i 的大小与他们的水平的高低是一致的, 故可用 $\bar{x} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_a)$ 来代替 \bar{x}^* 。

考虑到量 $|x_j - \bar{x}^*|$ 受运动员人数的影响, 及为了讨论方便起见, 我们用量

$$P_j = \frac{1}{a} \sum_{i=1}^a (x_{ij} - \bar{x}_i)^2$$

来表示第 j 号裁判员裁判水平的高低, P_j 越小, 水平越高, P_j 越大, 水平越低。

由上式计算可得出 P_j 结果, 见表 8-26。

表 8-26 平均偏差平方和计算表

裁判员	1	2	3	4
P_j	0.0315	0.0076786	0.00053571	0.0133299
名次	4	2	1	3

由表 8-26 可见, 3 号裁判员打分的平均偏差平方和最小, 因此, 在四位裁判中, 他的水平最高, 其他依次为 2 号、4 号, 1 号裁判员打分的平均偏差平方和最大, 相对而言, 他最差。

这里只是列举出用方差分析的方法来评估裁判员的水平, 显然, 这只是众多评价中的一种有效的方法, 但不是唯一的。

8.3.1.2 有缺失值的单因变量双因素无重复完全随机试验的方差分析

1. 缺失值估计的基本原理

在试验中, 由于各种各样的原因, 难免会出现某个试验方案下, 试验观测值缺失的情形, 在这种情况下, 必须要用统计方法估算出该缺失的数据值, 然后填进估值, 再进行统计分析, 否则行列的正交性遭到破坏。这是一种不得已的补救办法, 当缺失值较多时, 应重新安排试验。

缺失值估计的基本原理是最小平方法, 也就是使误差项的平方和为最小。

在上面已经提到, 试验资料的每一个观察值的线性可加模型可表示为

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (i=1,2,\dots,a; j=1,2,\dots,b)$$

当用样本估计时, 由最小平方法可得到上式中各参数的估计值, $\mu = \bar{x}$ 、 $\alpha_i = \bar{x}_i - \bar{x}$ 、 $\beta_j = \bar{x}_j - \bar{x}$ 和 $e_{ij} = x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x}$ 。因此, 对于任意一个观察值 x_{ij} , 它都可以表示为

$$x_{ij} = \bar{x}_i + \bar{x}_j - \bar{x}。$$

不失一般性, 假设缺失值在第 i 行、第 j 列, 它所在行的和为 T_i 、所在列的和为 T_c , 此时观察值的总和为 T , 则由上面的等式可知, 所要填入的估计值 x'_{ij} 应满足

$$x'_{ij} = \frac{T_i + x'_{ij}}{b} + \frac{T_c + x'_{ij}}{a} - \frac{T + x'_{ij}}{ab}$$

移项整理得

$$x'_{ij} = \frac{aT_i + bT_c - T}{(a-1)(b-1)}$$

由此, 可解得所要填入的 x'_{ij} 值。

当缺失值有两个时, 根据上式可得联立方程组, 来解得两个缺失值的估计值。

得到缺失值的估计值后,即可进行方差分析。同正常情况下的双因素方差分析时,所要改变的是一是总变异的自由度,二是随机误差的自由度,都要变小,在原有基础上再减去缺失值的个数。

因此,同前面的区别,除需要估计缺失值外,还要对正常情况下得到的方差分析表进行手工修正。

此外,有缺失值时,一般采用 LSD 法进行多重比较。

对于非缺失值的 A 因素各水平间进行比较时, $S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{2s_e^2}{b}}$, B 因素各水平间进行比较时, $S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{2s_e^2}{a}}$

对于 A 因素有缺失值水平与无缺失值的水平间进行比较时, $S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_e^2}{b} \left[2 + \frac{a}{(a-1)(b-1)} \right]}$, B 因素有缺失值水平与无缺失值的水平间进行比较时, $S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_e^2}{a} \left[2 + \frac{b}{(a-1)(b-1)} \right]}$

2. 实例分析

例 8.6 熟练工人操作四种不同的机器在六种不同运转速度加工某种零件,测得 24 小时内,不同运转速度下所生产的零件个数见表 8-27,其数据已经建成数据文件,存放在 data08-07.sav 中,试分析不同的机器和不同的运转速度对产量是否有影响?

表 8-27 24 小时生产的产量统计表

机器	运转速度					
	1	2	3	4	5	6
1	42.5	39.3	39.6		42.9	43.6
2	39.5	40.1	40.5	42.3	42.5	43.1
3	40.2	40.5	41.3	43.4	44.9	45.1
4	41.3	42.2	43.5	44.2	45.9	42.3

【题析】 由于本例数据中含有一个缺失值,因此,首先要计算该缺失值的估计值。

(1) 计算缺失值所在行和列的和及有缺失值状态下的观察值总和

在数据编辑窗口中,打开数据文件 data08-07.sav。

按 Analyze→Compare Means →Means 顺序,打开 Means 对话框,见图 8-8。选择零件数量和机器号变量移入到 Dependent List 下框中。

单击 Options 按钮,打开 Options 对话框,见图 8-9。在 Statistics 统计量框中选择 Sum 统

计量到 Cell Statistics 下框中, 在系统默认计算的统计量的基础上, 增加计算求和统计量。
单击 Continue 按钮, 返回 Means 对话框。

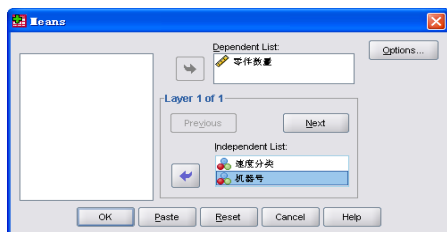


图 8-8 Means 对话框

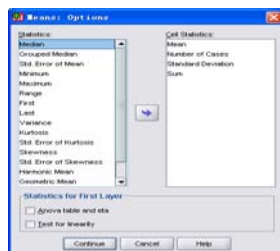


图 8-9 Means 的 Options 选项卡

单击 OK 按钮运行, 则在输出窗口中得到三张表, 其中后两张表为所需要的均值、求和表, 见表 8-28、表 8-29。

表 8-28 各速度水平下的均值和总和

零件数量 * 速度分类				
零件数量	Mean	N	Std. Deviation	Sum
速...	40.8750	4	1.31244	163.50
1	40.5250	4	1.22304	162.10
2	41.2250	4	1.66808	164.90
3	43.3000	3	.95394	129.90
4	44.0500	4	1.61967	176.20
5	43.5250	4	1.17863	174.10
6	42.2043	23	1.88595	970.70
Total				

表 8-29 各台机器的生产零件的均值和总和

零件数量 * 机器号				
零件数量	Mean	N	Std. Deviation	Sum
机...				
1	41.5800	5	1.98671	207.90
2	41.3333	6	1.48279	248.00
3	42.5667	6	2.19241	255.40
4	43.2333	6	1.66092	259.40
Total	42.2043	23	1.88595	970.70

(2) 计算缺失值的估计值

$$x'_{ij} = \frac{aT_i + bT_c - T}{(a-1)(b-1)} = \frac{207.90 \times 4 + 129.90 \times 6 - 970.70}{(4-1)(6-1)} = 42.68667$$

(3) 将缺失值补齐后计算方差分析

具体步骤同正常情况下的做法, 参见例 8.5。

可得方差分析表, 见表 8-30。

由于在有缺失值的情况下所做的方差分析, 要比正常情况下的双因素方差分析, 总变异的自由度和随机误差的自由度都要在原有基础上再减去缺失值的个数。因此, 校正的方差分析表见表 8-31, 表中手工修改部分用加粗的黑体字表示, 另外, sig 值是用 SPSS 中的函数重新计算得出。

从表 8-31 可见, 在不同机器间生产的零件数间无差异的原假设下, 出现目前统计量的值或者更极端值的概率为 0.061, 大于 0.05, 故现有的证据不足以拒绝原假设, 而在不同的操作速度下生产的零件数间无差异的原假设下, 出现目前统计量的值或者更极端值的概率为 0.002, 小于 0.01, 故拒绝原假设, 而认为不同的操作速度下生产的零件数的均值间有极显著性差异。

表 8-30 方差分析表

Tests of Between-Subjects Effects					
Source	Type IV Sum of Squares	df	Mean Square	F	Sig.
Model	42848.864 ^a	9	4760.985	3.701E3	.000
机器号	12.844	3	4.281	3.328	.048
速度分类	46.330	5	9.266	7.202	.001
Error	19.298	15	1.287		
Total	42868.162	24			

a. R Squared = 1.000 (Adjusted R Squared = .999)

表 8-31 校正的方差分析表

Tests of Between-Subjects Effects					
Source	Type IV Sum of Squares	df	Mean Square	F	Sig.
Model	42848.864 ^a	9	4760.985	3.454E3	.000
机器号	12.844	3	4.281	3.106	.061
速度分类	46.330	5	9.266	6.722	.002
Error	19.298	14	1.378		
Total	42868.162	23			

a. R Squared = 1.000 (Adjusted R Squared = .999)

(4) 对不同的操作速度下生产的零件数的均数间进行多重比较

首先, 用填入了缺失值估计值后的多重比较的结果。方法可参见例 8.1 解题步骤 7 中的相关步骤, 可得到表 8-32。

表 8-32 多重比较表

Multiple Comparisons									
因变量	对比	Mean Contrast	df	SS	df	SS	df	SS	df
零件数	0	Mean Contrast	df	SS	df	SS	df	SS	df
1	2	3500	88204	688	-1.3595	2.0595			
3	3	3500	88204	688	-2.0595	1.5595			
4	5	-2.2717	88204	013	-3.9812	-.9822			
5	6	-3.1750	88204	001	-4.8845	-1.4655			
6	0	2.8500	88204	005	-4.3595	-.8495			
2	1	-3500	88204	688	-2.0595	1.5595			
3	3	-7000	88204	297	-2.4595	1.6095			
4	5	-2.8217	88204	005	-4.3312	-.9122			
5	6	-3.5250	88204	001	-5.2345	-1.8155			
6	0	3.0800	88204	002	-4.7085	-1.2085			
3	1	3500	88204	688	-1.3595	2.0595			
2	3	7000	88204	297	-1.0895	2.4095			
4	5	-1.9217	88204	030	-3.6312	-.2122			
5	6	-2.8250	88204	003	-4.5345	-1.1155			
6	0	2.3800	88204	012	-4.0885	-.5885			
4	1	2.2717	88204	013	5822	3.9812			
2	3	2.8217	88204	005	9122	4.3312			
3	5	1.9217	88204	030	2122	3.6312			
5	6	-.9033	88204	278	-2.6128	8062			
6	0	-.3783	88204	644	-2.0878	1.3312			
5	1	3.1750	88204	001	1.4855	4.8845			
2	3	3.5250	88204	001	1.6155	5.2345			
4	6	2.8250	88204	003	1.1155	4.5345			
6	4	9033	88204	278	-.8062	2.6128			
0	6	5250	88204	523	1.1845	2.3445			
6	1	2.6500	88204	005	9495	4.3595			
2	3	3.0800	88204	002	1.2885	4.7085			
3	5	2.3000	88204	012	5885	4.0885			
4	6	-.3783	88204	644	-1.3312	2.0878			
0	5	-5250	88204	523	-2.2245	1.1845			

Based on observed means.
The error term is Mean Square(Error) = 1.287.

* The mean difference is significant at the 0.05 level.

	均值1	均值2	差异	均值差异的标准误
1	1	2	0.3500	0.83019
2	1	3	-0.3500	0.83019
3	1	4	-2.2717	0.90943
4	1	5	-3.1750	0.83019
5	1	6	-2.6500	0.83019
6	2	1	-0.3500	0.83019
7	2	3	-0.7000	0.83019
8	2	4	-2.6217	0.90943
9	2	5	-3.5250	0.83019
10	2	6	-3.0000	0.83019
11	3	1	0.3500	0.83019
12	3	2	0.7000	0.83019
13	3	4	-1.9217	0.90943
14	3	5	-2.8250	0.83019
15	3	6	-2.3000	0.83019
16	4	1	2.2717	0.90943
17	4	2	2.6217	0.90943
18	4	3	1.9217	0.90943
19	4	5	-0.9033	0.90943

图 8-10 整理后的多重比较基本数据

其次, 计算两两比较时均值差异的标准误。

由于 $s_e^2 = 1.37844$ (考虑到计算的精度, 因此它比表中显示的位数要多 2 位, 在输出窗口中双击该值, 即可看到它更多的有效位数), 因此, 对于非缺失值的操作速度因素各水平间进行比较时, 均值差异的标准误为

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{a}} = \sqrt{\frac{2 \times 1.37844}{4}} = 0.83019$$

对于操作速度因素有缺失值水平 (4) 与无缺失值的水平间进行比较时 (1、2、3、5、6), 均值差异的标准误为

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_e^2}{a} \left[2 + \frac{b}{(a-1)(b-1)} \right]} = \sqrt{\frac{1.37844}{4} \times \left(2 + \frac{6}{3 \times 5} \right)} = 0.90943059$$

由此, 可在 SPSS 的数据编辑窗口中得到以均值差异为基础数据的数据文件, 见图 8-10, 它存放在数据文件 data08-08.sav 中。

在本例中, 因为 $t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}} \sim t(14)$, 因此, 在 data08-08.sav 中, 按 Transform→Compute

表 8-33 多重比较结果

Case Summaries					
	均值1	均值2	差异	均值差异的标准误	概率
1	1	2	.3500	.83019	.68
2	1	3	-.3500	.83019	.68
3	1	4	-2.2717	.90943	.03
4	1	5	-3.1750	.83019	.00
5	1	6	-2.6500	.83019	.01
6	2	1	-.3500	.83019	.68
7	2	3	-.7000	.83019	.41
8	2	4	-2.6217	.90943	.01
9	2	5	-3.5250	.83019	.00
10	2	6	-3.0000	.83019	.00
11	3	1	.3500	.83019	.68
12	3	2	.7000	.83019	.41
13	3	4	-1.9217	.90943	.05
14	3	5	-2.8250	.83019	.00
15	3	6	-2.3000	.83019	.02
16	4	1	2.2717	.90943	.03
17	4	2	2.6217	.90943	.01
18	4	3	1.9217	.90943	.05
19	4	5	-.9033	.90943	.34
20	4	6	-.3783	.90943	.68
21	5	1	3.1750	.83019	.00
22	5	2	3.5250	.83019	.00
23	5	3	2.8250	.83019	.00
24	5	4	.9033	.90943	.34
25	5	6	.5250	.83019	.54
26	6	1	2.6500	.83019	.01
27	6	2	3.0000	.83019	.00
28	6	3	2.3000	.83019	.02
29	6	4	.3783	.90943	.68
30	6	5	-.5250	.83019	.54
Total	N	30	30	30	30

Variable 顺序, 展开 Compute Variable 对话框, 在 Target Variable 下框中输入目标变量名为“概率”, 在 Numeric Expression 下框中输入计算公式的表达式为“(1-CDF.T(abs(差异/均值差异的标准误),14))*2”, 则在数据编辑窗口出现新的一个变量“概率”及其计算结果。

按例 8.1 中最后结果的输出方法, 在输出窗口中可得表 8-33 所示的输出结果。

最后一列中概率小于 0.05 对应的两个均数之间有显著性差异。操作速度 1、2、3 之间生产的零件数均数之间无显著性差异, 操作速度 5、6、4 之间生产的零件数均数之间无显著性差异, 而操作速度 5、6、4 与操作速度 1、2、3 之间生产的零件数均数之间有显著性差异。

8.3.1.3 单因变量双因素有重复完全随机试验的方差分析

在双因素试验中, A 因素的每个水平和 B 因素的每个水平均衡相遇, 构成交叉分组, 当每个水平组合有多个重复观察值时, 这种资料就是单因变量有重复双因素完全随机试验资料。

此时, 除可以分析 A、B 因素对因变量的影响外, 还可讨论两个因素的联合作用, 也就是各因素的联合搭配对因变量的影响, 即两个因素之间的交互作用。

1. 有交互作用的双因素方差分析的基本原理

在两个因素无交互作用的情况下, $\mu_{ij} = \mu + \alpha_i + \beta_j, (i=1,2,\dots,a; j=1,2,\dots,b)$, 参数 μ 是 ab 个样本的总体平均值, 参数 α_i 表示因素 A 的各个不同水平的效应, 参数 β_j 表示因素 B 的各个不同水平的效应。

若 $\mu_{ij} \neq \mu + \alpha_i + \beta_j, (i=1,2,\dots,a; j=1,2,\dots,b)$, 则称 $\gamma_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j$ 为因素 A 的第 i 个水平与因素 B 的第 j 个水平的交互作用。 γ_{ij} 满足: $\sum_{i=1}^a \gamma_{ij} = 0, j=1,2,\dots,b$ 和

$$\sum_{j=1}^b \gamma_{ij} = 0, i=1, 2, \dots, a$$

2. 探讨交互作用时的实验重复次数及其实验结果记录

要探讨双因素之间有无交互作用，只凭一次试验结果是远远不够的，至少要做两次或两次以上的重复试验才能加以分析研究。不失一般性，假设为了研究两个因素之间的交互作用对试验指标是否有影响，在 (A_i, B_j) 组合下重复进行 n 次试验，对于因素 A_i 和 B_j 交互作用下的第 k 次实验的结果用 $y_{ijk} (k=1, 2, \dots, n)$ 表示，则有重复测定的实验结果记录参见表 8-34。

表 8-34 双因素重复测定的实验结果表

A	B				和	均值
	B_1	B_2	B_b		
A_1	y_{111}	y_{121}	y_{1b1}	$T_{1..}$	$\bar{y}_{1..}$
	y_{112}	y_{122}	y_{1b2}		
	\vdots	\vdots	\vdots		
	y_{11n}	y_{12n}	y_{1bn}		
A_2	y_{211}	y_{221}	y_{2b1}	$T_{2..}$	$\bar{y}_{2..}$
	y_{212}	y_{222}	y_{2b2}		
	\vdots	\vdots	\vdots		
	y_{21n}	y_{22n}	y_{2bn}		
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
A_i	y_{i11}	y_{i21}	y_{ib1}	$T_{i..}$	$\bar{y}_{i..}$
	y_{i12}	y_{i22}	y_{ib2}		
	\vdots	\vdots	\vdots		
	y_{i1n}	y_{i2n}	y_{ibn}		
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
A_a	y_{a11}	y_{a21}	y_{ab1}	$T_{a..}$	$\bar{y}_{a..}$
	y_{a12}	y_{a22}	y_{ab2}		
	\vdots	\vdots	\vdots		
	y_{a1n}	y_{a2n}	y_{abn}		
和	$T_{.1.}$	$T_{.2.}$	$T_{.b.}$	T	
均值	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	$\bar{y}_{.b.}$		\bar{y}

设同一交互作用下样本的和为 $T_{ij\bullet}$ ，同一交互作用下样本的均值为 $\bar{y}_{ij\bullet}$ ，则有

$$\begin{aligned} T_{ij\bullet} &= \sum_{k=1}^n y_{ijk} & \bar{y}_{ij\bullet} &= \frac{T_{ij\bullet}}{n} \\ T_{i\bullet\bullet} &= \sum_{j=1}^b \sum_{k=1}^n y_{ijk} & \bar{y}_{i\bullet\bullet} &= \frac{T_{i\bullet\bullet}}{bn} \\ T_{\bullet j\bullet} &= \sum_{i=1}^a \sum_{k=1}^n y_{ijk} & \bar{y}_{\bullet j\bullet} &= \frac{T_{\bullet j\bullet}}{an} \\ T &= \sum_{i=1}^a T_{i\bullet\bullet} = \sum_{j=1}^b T_{\bullet j\bullet} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n y_{ijk} & \bar{y} &= \frac{T}{abn} \end{aligned}$$

为后续计算方便，可将公式中涉及的中间计算量列表如下，见表 8-35。

表 8-35 中间计算量汇总

A \ B	B				
	B_1	B_2	...	B_b	和
A_1	$T_{11\bullet}$	$T_{12\bullet}$...	$T_{1b\bullet}$	$T_{1\bullet\bullet}$
A_2	$T_{21\bullet}$	$T_{22\bullet}$...	$T_{2b\bullet}$	$T_{2\bullet\bullet}$
\vdots	\vdots	\vdots	...	\vdots	\vdots
A_a	$T_{a1\bullet}$	$T_{a2\bullet}$...	$T_{ab\bullet}$	$T_{a\bullet\bullet}$
和	$T_{\bullet 1\bullet}$	$T_{\bullet 2\bullet}$...	$T_{\bullet b\bullet}$	T

3. 有交互作用的双因素方差分析的数学模型：

- (1) $y_{ijk} (i=1,2,\dots,a; j=1,2,\dots,b; k=1,2,\dots,n)$ 相互独立，分别服从 $N(\mu_{ijk}, \sigma^2)$ 分布；
- (2) $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} (i=1,2,\dots,a; j=1,2,\dots,b)$ 且

$$\sum_{i=1}^a \alpha_i = 0, \sum_{j=1}^b \beta_j = 0, \sum_{i=1}^a \gamma_{ij} = \sum_{j=1}^b \gamma_{ij} = 0$$

- (3) $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$ ，诸 ε_{ijk} 相互独立均服从 $N(0, \sigma^2)$ 分布。

4. 有交互作用的双因素方差分析的原假设为

$$\begin{aligned} H_{01} : \alpha_1 &= \alpha_2 = \dots = \alpha_a = 0 \\ H_{02} : \beta_1 &= \beta_2 = \dots = \beta_b = 0 \\ H_{03} : \gamma_{ij} &= 0, i=1,2,\dots,a; j=1,2,\dots,b \end{aligned}$$

5. 有交互作用的双因素方差分析的平方和分解原理

令

$$\begin{aligned}
 SS_T &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n \left(y_{ijk} - \bar{y} \right)^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n y_{ijk}^2 - \frac{T^2}{abn} \\
 SS_A &= bn \sum_{i=1}^a \left(\bar{y}_{i..} - \bar{y} \right)^2 = \frac{1}{bn} \sum_{i=1}^a T_{i..}^2 - \frac{T^2}{abn} \\
 SS_B &= an \sum_{j=1}^b \left(\bar{y}_{.j.} - \bar{y} \right)^2 = \frac{1}{an} \sum_{j=1}^b T_{.j.}^2 - \frac{T^2}{abn} \\
 SS_{A \times B} &= n \sum_{i=1}^a \sum_{j=1}^b \left(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y} \right)^2 \\
 &= \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b T_{ij.}^2 - \frac{1}{bn} \sum_{i=1}^a T_{i..}^2 - \frac{1}{an} \sum_{j=1}^b T_{.j.}^2 + \frac{T^2}{abn} \\
 SS_e &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n \left(y_{ijk} - \bar{y}_{ij.} \right)^2
 \end{aligned}$$

则与双因素方差分析相类似，总偏差平方和有如下恒等式

$$SS_T = SS_A + SS_B + SS_{A \times B} + SS_e$$

其中， SS_T 为总偏差平方和， SS_A, SS_B 分别为因素 A、因素 B 的主效应平方和， $SS_{A \times B}$ 为因素 A 与 B 交互作用的平方和， SS_e 为误差平方和，它是由随机波动引起的。

6. 期望均方

表 8-34 数据资料中各项变异的自由度和期望均方的组成见表 8-36。

表 8-36 单因变量有重复双因素完全随机试验资料自由度与期望均方

变异来源	自由度 (df)	期望均方		
		固定模型	随机模型	A 随机, B 固定
A	$a-1$	$bn\kappa_A^2 + \sigma_e^2$	$bn\sigma_A^2 + n\sigma_{A \times B}^2 + \sigma_e^2$	$bn\sigma_A^2 + \sigma_e^2$
B	$b-1$	$bn\kappa_B^2 + \sigma_e^2$	$bn\sigma_B^2 + n\sigma_{A \times B}^2 + \sigma_e^2$	$bn\sigma_B^2 + \sigma_e^2$
A × B	$(a-1)(b-1)$	$n\kappa_{A \times B}^2 + \sigma_e^2$	$n\sigma_{A \times B}^2 + \sigma_e^2$	$n\sigma_{A \times B}^2 + \sigma_e^2$
试验误差	$ab(n-1)$	σ_e^2	σ_e^2	σ_e^2
总变异	$abn-1$			

表中， a 表示 A 因素的水平数， b 表示 B 因素的水平数， σ_e^2 表示试验误差的方差， κ_A^2 表示固定模型中 A 因素效应方差， σ_A^2 表示随机模型中 A 因素效应方差， κ_B^2 表示固定模型中 B 因素效应方差， σ_B^2 表示随机模型中 B 因素效应方差。 $\kappa_{A \times B}^2$ 表示固定模型中 A × B

因素交互效应方差, $\sigma_{A \times B}^2$ 表示随机模型中 $A \times B$ 因素交互效应方差。

从表 8-36 可见, 不同的模型, 期望方差的组成成分也不相同, 故对因素效应和交互作用进行 F 检验的分母也不相同。

在固定模型和混合模型时, 均以误差项均方作分母; 在随机模型时, A、B 因素的效应检验以交互作用项的均方作分母, 交互效应的检验时, 以误差项均方作分母。

7. 方差分析表

根据以上的分析, 可以得到如表 8-37 所示的固定和混合模型下的方差分析表。在随机模型下, 因素 A 和 B 做 F 检验时公式中的分母为交互作用项的均方差。

表 8-37 双因素重 n 次重复测定实验结果的方差分析用表

方差来源	偏差平方和	自由度	均方	F
因素 A	SS_A	$a-1$	$S_A^2 = \frac{SS_A}{a-1}$	$F_1 = \frac{S_A^2}{S_e^2}$
因素 B	SS_B	$b-1$	$S_B^2 = \frac{SS_B}{b-1}$	$F_2 = \frac{S_B^2}{S_e^2}$
交互作用 $A \times B$	$SS_{A \times B}$	$(a-1)(b-1)$	$S_{A \times B}^2 = \frac{SS_{A \times B}}{(a-1)(b-1)}$	$F_3 = \frac{S_{A \times B}^2}{S_e^2}$
误差	SS_e	$ab(n-1)$	$S_e^2 = \frac{SS_e}{ab(n-1)}$	
总	SS_T	$abn-1$		

8. 多重比较

若用 LSD 法进行多重比较, 则当因素 A 各水平之间进行多重比较时, $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{bn}}$,

B 因素各水平间进行多重比较时, $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{an}}$, 两因素水平组合之间进行多重比较时,

$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{n}}$, t 分布的自由度为 $df = ab(n-1)$ 。

9. 实例分析

例 8.7 某篮球队为了训练弹跳能力, 用四种强度和四种密度搭配进行训练, 选取 32 名各方面条件基本相似的篮球运动员作为被试对象, 对各个训练组合随机安排 2 名运动员, 训练半年后, 测得各人的原地纵跳的提高值作为评价训练的效果, 结果见表 8-38 (单位 cm), 表中数据已存放在数据文件 data08-09 中, 试问强度、密度及强度和密度的交互作用是否对弹跳能力有影响?

表 8-38 四种强度和四种密度搭配进行训练半年后的纵跳提高值

密度 B 强度 A	B_1	B_2	B_3	B_4
A_1	6	7	10	12
	8	8	8	10
A_2	8	11	13	9
	10	9	12	9
A_3	11	14	9	9
	8	12	10	8
A_4	10	8	5	4
	8	7	6	4

【题析】显然这是一个双因素有重复试验的固定模型，可以讨论两因素的交互作用。在 SPSS 中的解题步骤如下：

(1) 在数据编辑窗口中，打开数据文件 data08-09.sav。

(2) 按 Analyze → General Linear Model → Univariate 顺序，打开 Univariate 主对话框，见图 8-1。

(3) 在左侧因素和协变量框中，选择纵跳提高值变量，单击向右箭头，将其送入 Dependent Variable 框；用同样方法分别将强度、密度二个因素变量送入 Fixed Factor (s) 框中。

(4) 单击 Model 按钮，打开 Model 对话框，见图 8-2。关闭 include intercept in model 选项。其他不作选择，采用系统默认的全因素模型，即在模型中既要分析 A、B 因素对因变量的影响，还要分析 A、B 因素的交互作用对因变量的影响，单击 Continue 按钮，返回 Univariate 主对话框。

(5) 单击 OK 按钮运行，则在输出窗中，得到两张表，其中第二张表为所需的两因素之间的方差分析表，见表 8-39。

(6) 结果与讨论

从表 8-39 可见，在强度因素各水平均值间无差异的原假设下，出现目前统计量的值或者更极端值的概率为

0.000，小于 0.05，故拒绝原假设；在密度因素各水平均值间无差异的原假设下，出现目前统计量的值或者更极端值的概率为 0.136，大于 0.05，故不拒绝原假设；在 A 因素和 B

表 8-39 方差分析表

Tests of Between-Subjects Effects					
Dependent Variable: 纵跳提高值					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	2661.500 ^a	16	166.344	123.791	.000
强度	70.594	3	23.531	17.512	.000
密度	8.594	3	2.865	2.132	.136
强度 * 密度	79.531	9	8.837	6.576	.001
Error	21.500	16	1.344		
Total	2683.000	32			

a. R Squared = .992 (Adjusted R Squared = .984)

因素的交互作用为 0 的原假设下, 出现目前统计量的值或者更极端值的概率为 0.001, 故拒绝原假设。

(7) 各因素的贡献率

因为

$$\begin{aligned}
 SS'_A &= SS_A - (a-1)S_e^2 = 70.59375 - 3 \times 1.344 = 66.56175 \\
 SS'_B &= SS_B - (b-1)S_e^2 = 8.59375 - 3 \times 1.344 = 4.56175 \\
 SS'_{A \times B} &= SS_{A \times B} - (a-1)(b-1)S_e^2 = 79.53125 - 9 \times 1.344 = 67.43525 \\
 SS'_e &= SS_e + (a-1)S_e^2 + (b-1)S_e^2 + (a-1)(b-1)S_e^2 \\
 &= 21.5 + 15 \times 1.344 \\
 &= 41.66
 \end{aligned}$$

所以

$$\begin{aligned}
 100\% &= \rho_A + \rho_B + \rho_{A \times B} + \rho_E \\
 &= \frac{66.56175}{180.21875} + \frac{4.56175}{180.21875} + \frac{67.43525}{180.21875} + \frac{41.66}{180.21875} \\
 &= 36.93\% + 2.53\% + 37.42\% + 23.12\%
 \end{aligned}$$

(8) 对 4 个不同强度进行多重比较

单击 Post Hoc 按钮, 展开 Post Hoc 多重比较对话框, 见图 8-6。在 Factor(s) 下框中, 将强度变量移入到 Post Hoc Tests for 下框中, 选择等方差假定前提下的 LSD 选项, 用 LSD 检验法进行多重比较。单击 Continue 按钮, 返回 Univariate 主对话框。

单击 OK 按钮运行, 则在输出窗中, 得到三张表, 第三张表为不同强度均值间的多重比较表, 见表 8-40。

表 8-40 不同强度均值间的多重比较

Multiple Comparisons					
LSD					
① 强度	② 强度	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval
1	2	-.15000 [*]	.57960	.020	-2.7287 - .2713
	3	-.15000 [*]	.57960	.020	-2.7287 - .2713
	4	2.1250 [*]	.57960	.002	.8963 3.3537
2	1	1.5000 [*]	.57960	.020	.2713 2.7287
	3	.0000	.57960	1.000	-1.2287 1.2287
	4	3.6250 [*]	.57960	.000	2.3963 4.8537
3	1	1.5000 [*]	.57960	.020	.2713 2.7287
	2	.0000	.57960	1.000	-1.2287 1.2287
	4	3.6250 [*]	.57960	.000	2.3963 4.8537
4	1	-2.1250 [*]	.57960	.002	-3.3537 -.8963
	2	-3.6250 [*]	.57960	.000	-4.8537 -2.3963
	3	-3.6250 [*]	.57960	.000	-4.8537 -2.3963

Based on observed means.
The error term is Mean Square(Error) = 1.344.
*. The mean difference is significant at the 0.05 level.

从表 8-40 可见, 4 个不同强度之间, 除强度 2 和 3 之间没有显著性差异外, 强度 2、3 和 1、4 之间, 强度 1 和 4 之间, 均有显著性差异。从均值上来看, 强度 2、3 对提高纵跳的效果最好, 强度 1 次之, 强度 4 最差。

(9) 交互作用分析

在上述前 4 步的基础上, 单击 Plots 按钮, 打开 Plots 对话框, 见图 8-11。将强度变量移入到 Horizontal Axis 下框中, 设定强度变量为输出图形的横轴, 将密度变量移入到 Separate

Lines 下框中, 要求输出各密度水平单独的线图。单击 Add 按钮, 则在 Plots 下框中出现所要的设置, 即要求输出强度和密度的交互作用图。单击 Continue 按钮, 返回 Univariate

主对话框。

单击 OK 按钮运行,则在输出窗中,在原有输出的基础上,增加一张所需的交互作用图,见图 8-12。图形的纵轴为纵跳提高值的均值。

从图 8-12 可见,4 条交互作用线之间彼此交叉,说明交互作用对纵跳提高值有影响。当强度为 3 水平、密度为 2 水平时,纵跳的提高值的均值最大,因此,此时为两个因素不同水平的最佳搭配。



图 8-11 Plots 对话框

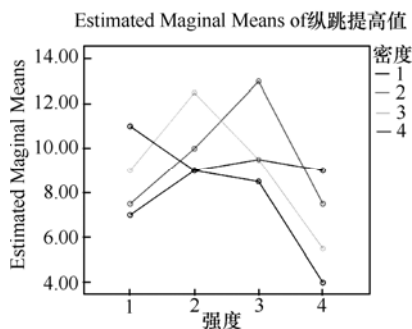


图 8-12 交互作用图

(10) 结论

由上分析可知,强度因素及其强度因素和密度因素的交互作用有显著性意义,密度因素各水平间无显著性差异。强度和密度的交互作用对总偏差平方和的贡献最大为 37.42%,其次为强度因素占总偏差平方和的 36.93%。残差占总偏差平方和的 23.12%,密度因素最小,仅限占总偏差平方和的 2.53%。这说明强度,尤其是强度和密度的有效搭配可以有效地提高弹跳能力。从原始的实验结果可知, A_3B_2 搭配效果最好。

8.3.1.4 单因变量双因素随机区组试验的方差分析

1. 数据模型

在双因素随机区组试验中,设 A 因素有 a 个水平, B 因素有 b 个水平,随机区组设计,有 n 次重复,则这样的试验中共有 abn 个观察值。设 x_{ijk} 为其中的任意一个观察值,则在该试验中获得的数据模型见表 8-41。

为后续计算方便,可将上表中涉及的中间计算量列表如下,见表 8-42。

在本模型下获取的数据资料观察值的线性模型可表示为

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_k + \varepsilon_{ijk}$$

$$(i = 1, 2, \dots, a; j = 1, 2, \dots, b; k = 1, 2, \dots, n)$$

其中, μ 为全部试验观察值的总体均值; α_i 为 A 因素第 i 个水平的效应; β_j 为 B 因素第 j 个水平的效应; $(\alpha\beta)_{ij}$ 为 A 因素第 i 个水平和 B 因素第 j 个水平的交互作用; δ_k 为第 k 个区组的效应; ε_{ijk} 为随机误差,相互独立且服从 $N(0, \sigma^2)$ 。

表 8-41 双因素随机区组试验资料的数据模型

A 因素	B 因素	区组				处理总和 T_i	处理均值 $\bar{x}_{i\bullet}$
		1	2	...	n		
A_1	B_1	x_{111}	x_{112}	...	x_{11n}	$T_{11\bullet}$	$\bar{x}_{11\bullet}$
	B_2	x_{121}	x_{122}	...	x_{12n}	$T_{12\bullet}$	$\bar{x}_{12\bullet}$
	\vdots	\vdots	\vdots	...	\vdots	\vdots	\vdots
	B_b	x_{1b1}	x_{1b2}	...	x_{1bn}	$T_{1b\bullet}$	$\bar{x}_{1b\bullet}$
A_2	B_1	x_{211}	x_{212}	...	x_{21n}	$T_{21\bullet}$	$\bar{x}_{21\bullet}$
	B_2	x_{221}	x_{222}	...	x_{22n}	$T_{22\bullet}$	$\bar{x}_{22\bullet}$
	\vdots	\vdots	\vdots	...	\vdots	\vdots	\vdots
	B_b	x_{2b1}	x_{2b2}	...	x_{2bn}	$T_{2b\bullet}$	$\bar{x}_{2b\bullet}$
\vdots	\vdots	\vdots	\vdots	...	\vdots	\vdots	\vdots
A_a	B_1	x_{a11}	x_{a12}	...	x_{a1n}	$T_{a1\bullet}$	$\bar{x}_{a1\bullet}$
	B_2	x_{a21}	x_{a22}	...	x_{a2n}	$T_{a2\bullet}$	$\bar{x}_{a2\bullet}$
	\vdots	\vdots	\vdots	...	\vdots	\vdots	\vdots
	B_b	x_{ab1}	x_{ab2}	...	x_{abn}	$T_{ab\bullet}$	$\bar{x}_{ab\bullet}$
区组总和		$T_{\bullet 1}$	$T_{\bullet 2}$...	$T_{\bullet n}$	T	\bar{x}

表 8-42 中间计算量汇总

A \ B	B				A 总和 T_A	A 均值 \bar{x}_A
	B_1	B_2	...	B_b		
A_1	$T_{11\bullet}$	$T_{12\bullet}$...	$T_{1b\bullet}$	$T_{1\bullet\bullet}$	$\bar{x}_{1\bullet\bullet}$
A_2	$T_{21\bullet}$	$T_{22\bullet}$...	$T_{2b\bullet}$	$T_{2\bullet\bullet}$	$\bar{x}_{2\bullet\bullet}$
\vdots	\vdots	\vdots	...	\vdots	\vdots	\vdots
A_a	$T_{a1\bullet}$	$T_{a2\bullet}$...	$T_{ab\bullet}$	$T_{a\bullet\bullet}$	$\bar{x}_{a\bullet\bullet}$
B 总和 T_B	$T_{\bullet 1}$	$T_{\bullet 2}$...	$T_{\bullet b}$	T	
B 均值 \bar{x}_b	$\bar{x}_{\bullet 1}$	$\bar{x}_{\bullet 2}$...	$\bar{x}_{\bullet b}$		\bar{x}

2. 有交互作用的双因素方差分析的原假设为

$$H_{01} : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

$$H_{02} : \beta_1 = \beta_2 = \cdots = \beta_b = 0$$

$$H_{03} : (\alpha\beta)_{ij} = 0, i=1,2,\cdots,a; j=1,2,\cdots,b$$

$$H_{04} : \delta_k = 0; k=1,2,\cdots,n$$

3. 双因素随机区组方差分析的平方和分解原理

令

$$\begin{aligned}
 \text{总偏差平方和: } SS_T &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (x_{ijk} - \bar{x})^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n x_{ijk}^2 - \frac{T^2}{abn} \\
 SS_A &= bn \sum_{i=1}^a (\bar{x}_{i..} - \bar{x})^2 = \frac{1}{bn} \sum_{i=1}^a T_{i..}^2 - \frac{T^2}{abn} \\
 SS_B &= an \sum_{j=1}^b (\bar{x}_{.j.} - \bar{x})^2 = \frac{1}{an} \sum_{j=1}^b T_{.j.}^2 - \frac{T^2}{abn} \\
 SS_r &= ab \sum_{k=1}^n (\bar{x}_{..k} - \bar{x})^2 = \frac{1}{ab} \sum_{k=1}^n T_{..k}^2 - \frac{T^2}{abn} \\
 SS_e &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (x_{ijk} - \bar{x}_{ij.})^2 - SS_r
 \end{aligned}$$

则与双因素方差分析相类似, 总偏差平方和有如下恒等式

$$SS_T = SS_A + SS_B + SS_{A \times B} + SS_r + SS_e$$

因此, $SS_{A \times B} = SS_T - SS_r - SS_e - SS_A - SS_B$

而处理: $SS_t = SS_T - SS_r - SS_e = SS_A + SS_B + SS_{A \times B}$

其中, SS_T 为总偏差平方和, SS_A, SS_B 分别为因素 A、因素 B 的主效应平方和, $SS_{A \times B}$ 为因素 A 与 B 交互作用的平方和, SS_r 为区组效应平方和, SS_e 为误差平方和, 它是由随机波动引起的。

4. 期望均方

表 8-41 数据资料中各项变异的自由度和期望均方的组成见表 8-43。

表 8-43 单因变量双因素随机区组试验资料自由度与期望均方

变异来源	自由度 (df)	期望均方		
		固定模型	随机模型	A 随机, B 固定
区组	$n-1$	$bn\kappa_r^2 + \sigma_e^2$	$bn\sigma_r^2 + \sigma_e^2$	
处理	$ab-1$			
A	$a-1$	$bn\kappa_A^2 + \sigma_e^2$	$bn\sigma_A^2 + n\sigma_{A \times B}^2 + \sigma_e^2$	$bn\sigma_A^2 + \sigma_e^2$
B	$b-1$	$bn\kappa_B^2 + \sigma_e^2$	$bn\sigma_B^2 + n\sigma_{A \times B}^2 + \sigma_e^2$	$bn\sigma_B^2 + \sigma_e^2$
A × B	$(a-1)(b-1)$	$n\kappa_{A \times B}^2 + \sigma_e^2$	$n\sigma_{A \times B}^2 + \sigma_e^2$	$n\sigma_{A \times B}^2 + \sigma_e^2$
试验误差	$(ab-1)(n-1)$	σ_e^2	σ_e^2	σ_e^2
总变异	$ab(n-1)$			

表中, a 表示 A 因素的水平数, b 表示 B 因素的水平数, n 表示区组数, σ_e^2 表示

试验误差的方差, κ_A^2 表示固定模型中 A 因素效应方差, κ_r^2 表示固定模型中区组效应方差, σ_r^2 表示随机模型中区组效应方差, σ_A^2 表示随机模型中 A 因素效应方差, κ_B^2 表示固定模型中 B 因素效应方差, σ_B^2 表示随机模型中 B 因素效应方差。 $\kappa_{A \times B}^2$ 表示固定模型中 $A \times B$ 因素交互效应方差, $\sigma_{A \times B}^2$ 表示随机模型中 $A \times B$ 因素交互效应方差。

从表 8-43 可见, 不同的模型, 期望方差的组成成分也不相同, 故对因素效应和交互作用进行 F 检验的分母也不尽相同。

在固定模型和混合模型时, 均以误差项均方作分母; 在随机模型时, A、B 因素的效应检验以交互作用项的均方作分母, 交互效应的检验时, 以误差项均方作分母。

5. 方差分析表

根据以上的分析, 可以得到如表 8-44 所示的固定和混合模型下的方差分析表。

表 8-44 双因素随机区组试验结果的方差分析用表

方差来源	偏差平方和	自由度	均方	F
区组	SS_r	$n-1$	$S_r^2 = \frac{SS_r}{n-1}$	$F_r = \frac{S_r^2}{S_e^2}$
处理	SS_t	$ab-1$	$S_t^2 = \frac{SS_t}{ab-1}$	$F_t = \frac{S_t^2}{S_e^2}$
因素 A	SS_A	$a-1$	$S_A^2 = \frac{SS_A}{a-1}$	$F_1 = \frac{S_A^2}{S_e^2}$
因素 B	SS_B	$b-1$	$S_B^2 = \frac{SS_B}{b-1}$	$F_2 = \frac{S_B^2}{S_e^2}$
交互作用 $A \times B$	$SS_{A \times B}$	$(a-1)(b-1)$	$S_{A \times B}^2 = \frac{SS_{A \times B}}{(a-1)(b-1)}$	$F_3 = \frac{S_{A \times B}^2}{S_e^2}$
误差	SS_e	$(ab-1)(n-1)$	$S_e^2 = \frac{SS_e}{(ab-1)(n-1)}$	
总	SS_T	$ab(n-1)$		

6. 多重比较

若用 LSD 法进行多重比较, 则当因素 A 各水平之间进行多重比较时,

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{bn}}, \text{ B 因素各水平间进行多重比较时, } S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{an}}$$

进行多重比较时, $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{n}}$, t 分布的自由度为 $df = (ab-1)(n-1)$ 。

7. 实例分析

例 8.8 玉米品种 (A) 与施肥 (B) 二因素随机区组试验, A 因素有 A_1 、 A_2 、 A_3 三个水平, B 因素有 B_1 、 B_2 、 B_3 三个水平, 重复 4 次, 小区计产面积 20m^2 , 田间排列和小区产量 ($\text{kg}/20\text{m}^2$) 如图 8-13 所示, 图中数据已按要求建成数据文件 data08-10.sav, 试做方差分析。(数据来源: 中国农业出版社 2007 年 8 月 金益《试验设计与统计分析》第一版 P197)

A_3B_2	A_1B_2	A_2B_1	A_2B_3	A_3B_3	A_2B_2	A_1B_3	A_3B_1	A_1B_1	区组 1
10.0	11.0	19.0	17.0	9.0	20.0	12.0	19.0	17.0	
A_2B_2	A_2B_1	A_2B_3	A_1B_2	A_1B_3	A_3B_2	A_1B_1	A_3B_3	A_3B_1	区组 2
19.0	13.0	16.0	14.0	8.0	8.0	15.0	8.0	18.0	
A_1B_3	A_3B_3	A_1B_2	A_3B_1	A_1B_1	A_2B_1	A_3B_2	A_2B_2	A_2B_3	区组 3
8.0	7.0	13.0	16.0	13.0	11.0	10.0	13.0	18.0	
A_3B_3	A_3B_1	A_2B_1	A_2B_2	A_3B_2	A_1B_3	A_2B_3	A_1B_1	A_1B_2	区组 4
7.0	17.0	11.0	14.0	8.0	9.0	16.0	14.0	12.0	

图 8-13 玉米品种与施肥随机区组试验田间排列和小区产量

在 SPSS 中的解题步骤如下:

(1) 在数据编辑窗口中, 打开数据文件 data08-10.sav。

(2) 按 Analyze → General Linear Model → Univariate 顺序, 打开 Univariate 主对话框, 见图 8-1。

(3) 在左侧因素和协变量框中, 选择产量变量, 单击向右箭头, 将其送入 Dependent Variable 框; 用同样方法分别将品种、施肥二个因素变量送入 Fixed Factor(s) 框中。选择区组变量并将其移入到 Random Factor[s]框中。

(4) 单击 Model 按钮, 打开 Model 对话框, 见图 8-2。在 Specify Model 选项中, 选择 Custom 选项, 要求对模型进行自定义。

在 Build Term(s) 的 Type 下拉列表中, 选择 Interaction 项, 要求在模型中考虑因素间的交互作用。

在左侧因素和协变量框中, 分别选择品种、施肥变量将其移入到 Model 框中, 同时选中品种和施肥变量将其移入到 Model 框中(要求求这两个因素的交互作用), 选择区组变量将其移入到 Model 框中。关闭 include intercept in model 选项, 其他保持系统默认选项, 单击 Continue 按钮, 返回 Univariate 主对话框。

(5) 单击 OK 按钮运行, 则在输出窗中, 得到三张表, 其中第二张表为所需的两因素之间的方差分析表, 见表 8-45。

如果要计算处理的偏差平方和等统计量, 可按表 8-44 所列的相应内容和计算处理偏

差平方和的公式： $SS_t = SS_T - SS_r - SS_e = SS_A + SS_B + SS_{A \times B}$ 来运算即可。由于它在分析中作用不大，本例略。

表 8-45 方差分析表

Tests of Between-Subjects Effects					
Dependent Variable: 产量					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
品种					
Hypothesis	118.389	2	59.194	20.167	.000
Error	70.444	24	2.935 ^a		
施肥					
Hypothesis	98.722	2	49.361	16.817	.000
Error	70.444	24	2.935 ^a		
品种 * 施肥					
Hypothesis	213.778	4	53.444	18.208	.000
Error	70.444	24	2.935 ^a		
区组					
Hypothesis	48.556	3	16.185	5.514	.005
Error	70.444	24	2.935 ^a		

a. R Squared = .666

(6) 结果与讨论

从表 8-45 可见，在品种因素、施肥因素、品种和施肥因素的交互作用对产量无效的原假设下，出现目前统计量的值或者更极端值的概率为 0.000，以及在区组因素对产量无效的原假设下，出现目前统计量的值或者更极端值的概率为 0.005，均小于 0.01，故拒绝原假设，而认为它们

均有极显著性意义。

(7) 多重比较

单击 Post Hoc 按钮，展开 Post Hoc 多重比较对话框，见图 8-6。在 Factor(s) 下框中，将品种和施肥变量移入到 Post Hoc Tests for 下框中，选择等方差假定前提下的 LSD 选项，用 LSD 检验法进行多重比较。单击 Continue 按钮，返回 Univariate 主对话框。

单击 OK 按钮运行，则在输出窗中，得到五张表，其中第四张表为不同品种均值间的多重比较表，见表 8-46，第五张表为不同施肥水平的均值间的多重比较表，见表 8-47。

表 8-46 不同品种均值间的多重比较

Multiple Comparisons						
产量 LSD						
① 品种	② 品种	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
A1	A2	-3.4167 [*]	.69943	.000	-4.8602	-1.9731
	A3	.7500	.69943	.294	-.6935	2.1935
A2	A1	3.4167 [*]	.69943	.000	1.9731	4.8602
	A3	4.1667 [*]	.69943	.000	2.7231	5.6102
A3	A1	-.7500	.69943	.294	-2.1935	.6935
	A2	-4.1667 [*]	.69943	.000	-5.6102	-2.7231

Based on observed means.

The error term is Mean Square(Error) = 2.935.

*. The mean difference is significant at the 0.05 level.

表 8-47 不同施肥水平的均值间的多重比较

Multiple Comparisons						
产量 LSD						
① 施肥	② 施肥	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
B1	B2	2.5833 [*]	.69943	.001	1.1398	4.0269
	B3	4.0000 [*]	.69943	.000	2.5565	5.4435
B2	B1	-2.5833 [*]	.69943	.001	-4.0269	-1.1398
	B3	1.4167	.69943	.054	-.0269	2.8602
B3	B1	-4.0000 [*]	.69943	.000	-5.4435	-2.5565
	B2	-1.4167	.69943	.054	-2.8602	.0269

Based on observed means.

The error term is Mean Square(Error) = 2.935.

*. The mean difference is significant at the 0.05 level.

从表 8-46 可见，品种 A_2 均值最高，它与 A_1 和 A_3 之间均有极显著性差异 ($P=0.000<0.01$)，而品种 A_1 和 A_3 之间均无显著性差异 ($P=0.294>0.05$)。

从表 8-47 可见，不同施肥水平之间， B_1 的平均产量最高，它与 B_2 和 B_3 之间有极显著性差异 ($P=0.001$ 和 $0.000<0.01$)，而 B_2 和 B_3 之间无显著性差异 ($P=0.054>0.05$)。

(8) 交互作用分析

在上述前 4 步的基础上，单击 Plots 按钮，打开 Plots 对话框，见图 8-11。将品种变

量移入到 Horizontal Axis 下框中, 设定品种变量为输出图形的横轴, 将施肥变量移入到 Separate Lines 下框中, 要求输出各个施肥水平单独的线图。单击 Add 按钮, 则在 Plots 下框中出现所要的设置, 即要求输出不同品种和不同施肥水平间的交互作用图。单击 Continue 按钮, 返回 Univariate 主对话框。

单击 OK 按钮运行, 则在输出窗中, 在原有输出的基础上, 增加一张所需的交互作用图, 见图 8-14。图形的纵轴为产量的均值。

从图 8-14 可见, 3 条交互作用线之间彼此交叉, 说明不同品种和施肥不同水平间的交互作用对产量有影响。当品种为 A_2 时、以 B_3 、 B_2 施肥量为宜, 当品种为 A_1 、 A_3 时, 以 B_1 施肥量为宜, 在 A_3B_1 搭配时, 产量均值达到最大, 也就是它们是现有水平中的最佳搭配方案。

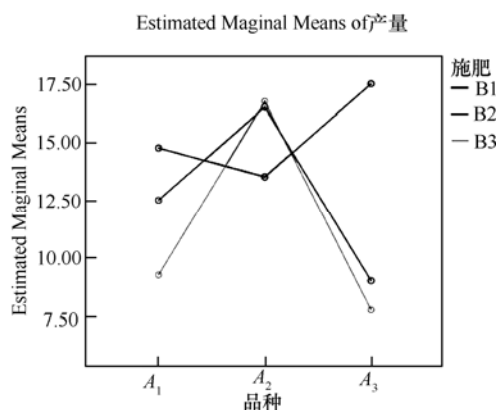


图 8-14 品种与施肥的交互作用图

(9) 结论

由上分析可知, 区组间有显著性差异, 说明土壤的肥力有差异, 品种因素、施肥因素及其它们间的交互作用有显著性意义。在品种因素方面, 品种 A_2 均值最高, 在施肥因素方面, B_1 对应的平均产量最高, A_2 品种应以 B_3 、 B_2 施肥量为宜, A_1 、 A_3 品种应以 B_1 施肥量为宜, 综合交互作用来看, 采用 A_3B_1 搭配时可获得产量的最大值。

8.3.1.5 单因变量双因素裂区试验的方差分析

裂区设计的本质是一种二次随机区组(或拉丁方)设计。设 A、B 两个试验因素, 分别有 a 和 b 个水平。首先对 A 因素 a 个水平进行随机区组设计, 然后将 A 因素的各个水平的试验单元再分成 b 个较小的试验单元, 再按随机区组设计将 B 因素的各水平安排入 A 因素的每个水平的每一个试验单元中。称安排 A 因素各个水平的试验单元为主区, 在主区内所安排的处理称为主处理; 称安排 B 因素各个水平的试验单元为副区或裂区, 在副区内所安排的处理称为副处理。

双因素裂区设计常用于农业科学研究的田间试验中, 它将两个因素分为主区因素(A)和副区因素(B)后, 分别进行安排的试验设计方法。裂区设计的示意图参见图 8-15。现也被广泛地应用到医学研究和其他领域的研究中。

在方差分析时, 要分别估计出主区误差和副区误差, 并按主区部分和副区部分分别进行分析。

也许细心的读者已经发现, 在 8.2 中介绍的单因变量单因素随机区组设计的方差分析

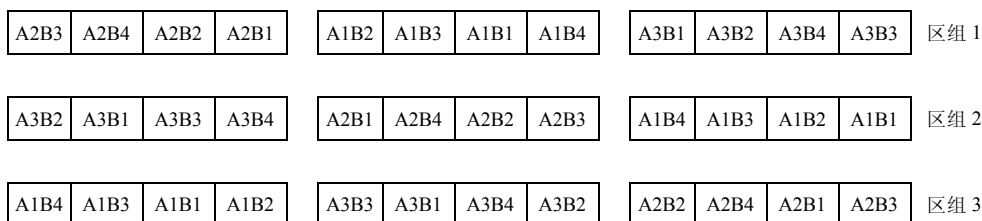


图 8-15 A 因素 3 水平和 B 因素 4 水平重复 3 次的裂区设计示意图

方法和 8.3.1.1 中介绍的单因变量双因素无重复试验设计的方差分析的方法是无本质上的区别的，主要区别是在对因素的定义上，定义为单因素还是双因素主要是看试验安排的因素，而不是影响试验结果的因素，这种区分不是没有道理的，虽然，仅就上面的情形，似乎无须区分，都能得到相同的结果，但我们很可能惯性思维到双因素随机区组试验和双因素裂区试验同三因素随机试验也是一样的情形，这就可能要进入误区了。在本节见到的双因素裂区试验中，如果按下一部分见到的三因素来处理，就会得到不一样的结果，显然，它同下一节要见到的三因素完全随机试验的方差分析在偏差平方和的分解上是有区别的，参见其观察值的线性模型，在 SPSS 中处理的一些细节上也有细微的区别，很容易混淆，故单列出来进行介绍，目的就是为了防止误用。

1. 数据模型

在双因素随机裂区设计试验中，设主区因素 A 有 a 个水平，副区因素 B 有 b 个水平，裂区设计，有 n 次重复，则这样的试验中共有 abn 个观察值。设 A_i 、 B_j 水平组合在第 k 个区组的观察值为 x_{ijk} ，则在该试验中获得的数据模型见表 8-48。

表 8-48 双因素随机区组试验资料的数据模型

主处理 A	副处理 B	区组				处理总和 T_i	处理均值 $\bar{x}_{i\cdot}$
		1	2	...	n		
A_1	B_1	x_{111}	x_{112}	...	x_{11n}	$T_{1\cdot}$	$\bar{x}_{1\cdot}$
	B_2	x_{121}	x_{122}	...	x_{12n}	$T_{12\cdot}$	$\bar{x}_{12\cdot}$
	\vdots	\vdots	\vdots	...	\vdots	\vdots	\vdots
	B_b	x_{1b1}	x_{1b2}	...	x_{1bn}	$T_{1b\cdot}$	$\bar{x}_{1b\cdot}$
	主区总和 $T_{1\cdot}$	$T_{1\cdot 1}$	$T_{1\cdot 2}$...	$T_{1\cdot n}$		
A_2	B_1	x_{211}	x_{212}	...	x_{21n}	$T_{21\cdot}$	$\bar{x}_{21\cdot}$
	B_2	x_{221}	x_{222}	...	x_{22n}	$T_{22\cdot}$	$\bar{x}_{22\cdot}$
	\vdots	\vdots	\vdots	...	\vdots	\vdots	\vdots
	B_b	x_{2b1}	x_{2b2}	...	x_{2bn}	$T_{2b\cdot}$	$\bar{x}_{2b\cdot}$

续表

主处理 A	副处理 B	区组				处理总和 T_i	处理均值 $\bar{x}_{i\bullet}$
		1	2	...	n		
	主区总和 T_m	$T_{2\bullet 1}$	$T_{2\bullet 2}$...	$T_{2\bullet n}$		
\vdots	\vdots	\vdots	\vdots	...	\vdots	\vdots	\vdots
A_a	B_1	x_{a11}	x_{a12}	...	x_{a1n}	$T_{a1\bullet}$	$\bar{x}_{a1\bullet}$
	B_2	x_{a21}	x_{a22}	...	x_{a2n}	$T_{a2\bullet}$	$\bar{x}_{a2\bullet}$
	\vdots	\vdots	\vdots	...	\vdots	\vdots	\vdots
	B_b	x_{ab1}	x_{ab2}	...	x_{abn}	$T_{ab\bullet}$	$\bar{x}_{ab\bullet}$
	主区总和 T_m	$T_{a\bullet 1}$	$T_{a\bullet 2}$...	$T_{a\bullet n}$		
区组总和 T_r		$T_{\bullet\bullet 1}$	$T_{\bullet\bullet 2}$...	$T_{\bullet\bullet n}$	T	\bar{x}

为后续计算方便，可将上表中涉及的中间计算量列表如下，见表 8-49。

表 8-49 中间计算量汇总

A \ B						
	B_1	B_2	...	B_b	A 总和 T_A	A 均值 \bar{x}_A
A_1	$T_{11\bullet}$	$T_{12\bullet}$...	$T_{1b\bullet}$	$T_{1\bullet\bullet}$	$\bar{x}_{1\bullet\bullet}$
A_2	$T_{21\bullet}$	$T_{22\bullet}$...	$T_{2b\bullet}$	$T_{2\bullet\bullet}$	$\bar{x}_{2\bullet\bullet}$
\vdots	\vdots	\vdots	...	\vdots	\vdots	
A_a	$T_{a1\bullet}$	$T_{a2\bullet}$...	$T_{ab\bullet}$	$T_{a\bullet\bullet}$	$\bar{x}_{a\bullet\bullet}$
B 总和 T_B	$T_{\bullet 1\bullet}$	$T_{\bullet 2\bullet}$...	$T_{\bullet b\bullet}$	T	
B 均值 \bar{x}_b	$\bar{x}_{\bullet 1\bullet}$	$\bar{x}_{\bullet 2\bullet}$...	$\bar{x}_{\bullet b\bullet}$		\bar{x}

在本模型下获取的数据资料观察值的线性模型可表示为

$$x_{ijk} = \mu + \delta_k + \alpha_i + (\varepsilon_a)_{ik} + \beta_j + (\alpha\beta)_{ij} + (\varepsilon_b)_{ijk}$$

$$(i=1,2,\dots,a; j=1,2,\dots,b; k=1,2,\dots,n)$$

其中， μ 为全部试验观察值的总体均值； α_i 为主区因素 A 第 i 个水平的效应； β_j 为副区 B 因素第 j 个水平的效应； $(\alpha\beta)_{ij}$ 为 A 因素第 i 个水平和 B 因素第 j 个水平的交互作用； δ_k 为第 k 个区组的效应； $(\varepsilon_a)_{ik}$ 和 $(\varepsilon_b)_{ijk}$ 分别为主区和副区误差，两者各自独立且分别服

从 $N(0, \sigma_{E_a}^2)$ 和 $N(0, \sigma_{E_b}^2)$ 。

2. 双因素裂区试验方差分析的原假设为

$$H_{01}: \delta_k = 0, k = 1, 2, \dots, n$$

$$H_{02}: \alpha_1 = \alpha_2 = \dots = \alpha_a = 0$$

$$H_{03}: \beta_1 = \beta_2 = \dots = \beta_b = 0$$

$$H_{04}: (\alpha\beta)_{ij} = 0, i = 1, 2, \dots, a; j = 1, 2, \dots, b$$

3. 双因素裂区试验方差分析的平方和分解原理

令

$$SS_T = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (x_{ijk} - \bar{x})^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n x_{ijk}^2 - \frac{T^2}{abn}$$

$$SS_m = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (\bar{x}_{i\cdot k} - \bar{x})^2 = \frac{\sum_{i=1}^a \sum_{k=1}^n T_{i\cdot k}^2}{b} - \frac{T^2}{abn} \quad (\text{主区总偏差平方和})$$

$$SS_r = ab \sum_{k=1}^n (\bar{x}_{\cdot\cdot k} - \bar{x})^2 = \frac{1}{ab} \sum_{k=1}^n T_{\cdot\cdot k}^2 - \frac{T^2}{abn}$$

$$SS_A = bn \sum_{i=1}^a (\bar{x}_{i\cdot\cdot} - \bar{x})^2 = \frac{1}{bn} \sum_{i=1}^a T_{i\cdot\cdot}^2 - \frac{T^2}{abn}$$

则

$$SS_{E_a} = SS_m - SS_r - SS_A$$

令

$$SS_t = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (\bar{x}_{ij\cdot} - \bar{x})^2 = \frac{\sum_{i=1}^a \sum_{j=1}^b T_{ij\cdot}^2}{n} - \frac{T^2}{abn}$$

$$SS_B = an \sum_{j=1}^b (\bar{x}_{\cdot j\cdot} - \bar{x})^2 = \frac{1}{an} \sum_{j=1}^b T_{\cdot j\cdot}^2 - \frac{T^2}{abn}$$

则

$$SS_{A \times B} = SS_t - SS_A - SS_B$$

$$SS_{E_b} = SS_T - SS_m - SS_B - SS_{A \times B}$$

与双因素随机区组设计方差分析相类似，总偏差平方和有如下恒等式

$$SS_T = SS_r + SS_A + SS_{E_a} + SS_B + SS_{A \times B} + SS_{E_b}$$

其中, SS_T 为总偏差平方和, SS_r 为区组效应平方和, SS_A 为因素 A 主效应平方和, SS_{E_a} 为主区误差平方和, 它是由随机波动引起的, SS_B 为因素 B 的主效应平方和, $SS_{A \times B}$ 为因素 A 与 B 交互作用的平方和, SS_{E_b} 为副区误差平方和, 它是由随机波动引起的。

4. 期望均方

表 8-48 数据资料中各项变异的自由度和期望均方的组成见表 8-50。

表 8-50 单因变量双因素裂区试验资料自由度与期望均方

变异来源		自由度 (df)	期望均方			
			固定模型	随机模型	A 随机, B 固定	A 固定, B 随机
主 区 部 分	区组	$n-1$	$\sigma_{E_b}^2 + b\sigma_{E_a}^2 + ab\kappa_r^2$	$\sigma_{E_b}^2 + b\sigma_{E_a}^2 + ab\sigma_r^2$		
	A	$a-1$	$\sigma_{E_b}^2 + b\sigma_{E_a}^2 + bn\kappa_A^2$	$\sigma_{E_b}^2 + b\sigma_{E_a}^2 + n\sigma_{A \times B}^2 + bn\sigma_A^2$	$\sigma_{E_b}^2 + b\sigma_{E_a}^2 + bn\kappa_A^2$	$\sigma_{E_b}^2 + b\sigma_{E_a}^2 + n\sigma_{A \times B}^2 + bn\kappa_A^2$
	E_a	$(n-1)(a-1)$	$\sigma_{E_b}^2 + b\sigma_{E_a}^2$	$\sigma_{E_b}^2 + b\sigma_{E_a}^2$	$\sigma_{E_b}^2 + b\sigma_{E_a}^2$	$\sigma_{E_b}^2 + b\sigma_{E_a}^2$
副 区 部 分	B	$b-1$	$\sigma_{E_b}^2 + an\kappa_B^2$	$\sigma_{E_b}^2 + n\sigma_{A \times B}^2 + an\sigma_B^2$	$\sigma_{E_b}^2 + n\sigma_{A \times B}^2 + an\kappa_B^2$	$\sigma_{E_b}^2 + an\sigma_B^2$
	A × B	$(a-1)(b-1)$	$\sigma_{E_b}^2 + n\kappa_{A \times B}^2$	$\sigma_{E_b}^2 + n\sigma_{A \times B}^2$	$\sigma_{E_b}^2 + n\sigma_{A \times B}^2$	$\sigma_{E_b}^2 + n\sigma_{A \times B}^2$
	E_b	$a(b-1)(n-1)$	$\sigma_{E_b}^2$	$\sigma_{E_b}^2$	$\sigma_{E_b}^2$	$\sigma_{E_b}^2$
总变异		$abn-1$				

表中, a 表示主区 A 因素的水平数, b 表示副区 B 因素的水平数, n 表示区组数, $\sigma_{E_a}^2$ 表示主区部分试验误差的方差, $\sigma_{E_b}^2$ 表示副区部分试验误差的方差, κ_A^2 表示固定模型中 A 因素效应方差, κ_r^2 表示固定模型中区组效应方差, σ_r^2 表示随机模型中区组效应方差, σ_A^2 表示随机模型中 A 因素效应方差, κ_B^2 表示固定模型中 B 因素效应方差, σ_B^2 表示随机模型中 B 因素效应方差, $\kappa_{A \times B}^2$ 表示固定模型中 A × B 因素交互效应方差, $\sigma_{A \times B}^2$ 表示随机模型中 A × B 因素交互效应方差。

在进行 F 检验时, F 计算公式中的分子和分母只能相差一个分量, 因此, 根据表 8-49, 即可构造不同的模型下的 F 检验的公式。

除对于主区因素 A 在随机模型和 A 固定、B 随机模型下进行方差分析时, 需要重新合并各项误差以此来构造 F 检验外, 其余可依表 8-50 得到 F 检验公式。

例如, 在固定模型 (I) 和 A 随机、B 固定模型 (II) 下, 对主区效应和副区效应进

行方差分析时, F 值的计算可参见表 8-51。

对应副区因素 B 在随机模型下的 F 检验公式同 A 随机、B 固定模型, 而 A 固定、B 随机模型下的检验公式同固定模型。

表 8-51 双因素裂区试验设计结果在固定模型和 A 随机、B 固定模型下的方差分析用表

方差来源		偏差平方和	自由度	均方	I	II
					F	F
主 区 部 分	区组	SS_r	$n-1$	$S_r^2 = \frac{SS_r}{n-1}$	$F_r = \frac{S_r^2}{S_{E_a}^2}$	$F_r = \frac{S_r^2}{S_{E_a}^2}$
	因素 A	SS_A	$a-1$	$S_A^2 = \frac{SS_A}{a-1}$	$F_1 = \frac{S_A^2}{S_{E_a}^2}$	$F_2 = \frac{S_A^2}{S_{A \times B}^2}$
	E_a	SS_E	$(n-1)(a-1)$	$S_{E_a}^2 = \frac{SS_{E_a}}{(n-1)(a-1)}$		
副 区 部 分	因素 B	SS_B	$b-1$	$S_B^2 = \frac{SS_B}{b-1}$	$F_2 = \frac{S_B^2}{S_{E_b}^2}$	$F_2 = \frac{S_B^2}{S_{A \times B}^2}$
	交互作用 $A \times B$	$SS_{A \times B}$	$(a-1)(b-1)$	$S_{A \times B}^2 = \frac{SS_{A \times B}}{(a-1)(b-1)}$	$F_3 = \frac{S_{A \times B}^2}{S_{E_b}^2}$	$F_3 = \frac{S_{A \times B}^2}{S_{E_b}^2}$
	E_b	SS_{E_b}	$a(b-1)(n-1)$	$S_e^2 = \frac{SS_{E_b}}{a(b-1)(n-1)}$		
总变异		SS_T	$ab(n-1)$			

在随机模型和 A 固定、B 随机模型下对主区因素 A 进行方差分析时, 可用 $F = \frac{S_1^2}{S_2^2}$ 进行, 其中, 用 $S_1^2 = S_A^2 + S_{E_b}^2$ 估计 $2\sigma_{E_b}^2 + b\sigma_{E_a}^2 + n\sigma_{A \times B}^2 + bn\sigma_A^2$ (或 $bn\kappa_A^2$), 而用 $S_2^2 = S_{A \times B}^2 + S_{E_b}^2$ 估计 $2\sigma_{E_b}^2 + b\sigma_{E_a}^2 + n\sigma_{A \times B}^2$, 此时近似的自由度为

$$df_1 = \frac{S_1^2}{\frac{S_A^2}{df_A} + \frac{S_{E_b}^2}{df_{E_b}}}, \quad df_2 = \frac{S_2^2}{\frac{S_{A \times B}^2}{df_{A \times B}} + \frac{S_{E_b}^2}{df_{E_b}}}$$

当交互作用的 F 检验不显著时, 可以不用上面近似的 F 检验, 而直接改用在固定模型下的主区因素 A 的 F 检验方法进行。

5. 多重比较

若用 LSD 法进行多重比较, 在固定模型下, 对主区因素各水平之间进行多重比较时,

$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_{E_a}^2}{bn}}$, 自由度为 $(n-1)(a-1)$; 对副区因素 B 各水平间和交互作用各水平间进

行多重比较时, 要用 $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_{E_b}^2}{an}}$, 自由度为 $a(b-1)(n-1)$ 。

6. 实例分析

例 8.9 有一水稻施 N 肥 (A) 与品种 (B) 试验, 主处理为 A, 有 A_1 、 A_2 、 A_3 共 3 个水平; 副处理 B 有 B_1 、 B_2 、 B_3 、 B_4 4 个水平; 裂区设计, 重复 3 次, 副区计产面积 13.34 m^2 , 试验处理与代号、田间排列和小区产量 (kg) 见图 8-16, 图中数据已按要求建成数据文件 data08-11.sav, 试做方差分析。(数据来源: 中国农业出版社 2007 年 8 月 金益《试验设计与统计分析》第一版 P203)

A3 B2 17	A3 B1 16	A3 B3 21	A3 B4 20	A2 B1 19	A2 B4 25	A2 B2 20	A2 B3 19	A1 B4 15	A1 B3 11	A1 B2 13	A1 B1 24
A1 B4 18	A1 B3 10	A1 B1 14	A1 B2 12	A3 B3 25	A3 B1 18	A3 B4 19	A3 B2 16	A2 B2 22	A2 B4 26	A2 B1 21	A2 B3 23
A2 B3 24	A2 B4 23	A2 B2 21	A2 B1 22	A1 B2 12	A1 B3 11	A1 B1 13	A1 B4 19	A3 B1 19	A3 B2 20	A3 B4 21	A3 B3 27

图 8-16 水稻施 N 肥与品种裂区试验田间排列和小区产量

在 SPSS 中的解题步骤如下:

(1) 在数据编辑窗口中, 打开数据文件 data08-11.sav。

(2) 按 Analyze → General Linear Model → Univariate 顺序, 打开 Univariate 主对话框, 见图 8-1。

(3) 在左侧因素和协变量框中, 选择产量变量, 单击向右箭头, 将其送入 Dependent Variable 框; 用同样方法分别将施肥、品种二个因素变量送入 Fixed Factor(s) 框中。选择区组变量并将其移入到 Random Factor[s] 框中。

(4) 单击 Model 按钮, 打开 Model 对话框, 见图 8-2。在 Specify Model 选项中, 选择 Custom 选项, 要求对模型进行自定义。

在 Build Term(s) 的 Type 下拉式选项卡中, 选择 Interaction 项, 要求在模型中考虑两个因素的交互作用。

在左侧因素和协变量框中, 按以下顺序分别选择区组、施肥、区组*施肥、品种、施肥*品种变量将其移入到 Model 框中。关闭 include intercept in model 选项, 其他保持系统默认选项, 单击 Continue 按钮, 返回 Univariate 主对话框。

注: 模型中要求计算的区组*施肥的交互作用项, 它实际上就是主区误差项。这是解

本类型题时需要注意的。否则，得不到正确的结果。

(5) 单击 OK 按钮运行，则在输出窗中，得到三张表，其中第二张表为所需的两因素之间的方差分析表，见表 8-52。

表 8-52 方差分析表

Tests of Between-Subjects Effects					
Dependent Variable: 产量					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
区组					
Hypothesis	20.667	2	10.333	3.139	.151
Error	13.167	4	3.292 ^a		
施肥					
Hypothesis	478.167	2	239.083	72.633	.001
Error	13.167	4	3.292 ^a		
施肥 * 区组					
Hypothesis	13.167	4	3.292	1.638	.208
Error	36.167	18	2.009 ^a		
品种					
Hypothesis	77.000	3	25.667	12.774	.000
Error	36.167	18	2.009 ^a		
施肥 * 品种					
Hypothesis	113.833	6	18.972	9.442	.000
Error	36.167	18	2.009 ^a		

a. MS(施肥 * 区组)

b. MS(Fertilizer)

(6) 结果与讨论

从表 8-52 可见，主区因素 A 的各水平（各种施 N 肥）间（ $P=0.001<0.01$ ）、副区因素 B 的各水平（不同品种）间（ $P=0.000<0.01$ ）的差异，A、B 两因素的交互作用（ $P=0.000<0.01$ ），均有极显著性意义。

(7) 多重比较

主区因素 A 不同水平间的多重比较，不能直接从现成程序的输出中得到，应按例 8.1 嵌套设计中的做法来完成。本例中，只给出结果，见表 8-53。

副区因素 B 不同水平间的多重比较，可单击 Post Hoc 按钮，展开 Post Hoc 多重比较对话框，见图 8-6。在 Factor(s) 下框中，将品种变量移入到 Post Hoc Tests for 下框中，选择等方差假定前提下的 LSD 选项[本例中，无法对方差是否齐性作出检验，而如果假定方差不齐，则会得到各品种的平均产量间均无显著差异的统计结论，这与上面得出的结论（不同品种间的差异显著）相悖，因此，假定方差齐性]，用 LSD 检验法进行多重比较。单击 Continue 按钮返回 Univariate 主对话框。

单击 OK 按钮运行，则在输出窗中，得到四张表，其中第四张表为不同品种均值间的多重比较表，见表 8-54。

表 8-53 不同施 N 肥均值间的多重比较

Case Summaries					
	施N肥I	施N肥J	均值差	标准误	概率
1	A1	A2	-8.5833	.7405	.0003
2	A1	A3	-6.4167	.7405	.0010
3	A2	A1	8.5833	.7405	.0003
4	A2	A3	2.1667	.7405	.0430
5	A3	A1	6.4167	.7405	.0010
6	A3	A2	-2.1667	.7405	.0430

表 8-54 不同品种均值间的多重比较

Multiple Comparisons						
产量 LSD						
因 子	因 子	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
B1	B2	.3333	.66821	.624	-1.0705	1.7372
	B3	-1.6667	.66821	.023	-3.0705	-.2628
	B4	-3.3333	.66821	.000	-4.7372	-1.9295
B2	B1	-.3333	.66821	.624	-1.7372	1.0705
	B3	-2.0000	.66821	.008	-3.4039	-.5961
	B4	-3.6667	.66821	.000	-5.0705	-2.2628
B3	B1	1.6667	.66821	.023	.2628	3.0705
	B2	2.0000	.66821	.008	.5961	3.4039
	B4	-1.6667	.66821	.023	-3.0705	-.2628
B4	B1	3.3333	.66821	.000	1.9295	4.7372
	B2	3.6667	.66821	.000	2.2628	5.0705
	B3	1.6667	.66821	.023	.2628	3.0705

Based on observed means.

The error term is Mean Square(Error) = 2.009.

*. The mean difference is significant at the .05 level.

从表 8-53 可见，施 N 肥 A_2 时均值最高，它与 A_1 和 A_3 之间均有显著性差异（ $P=0.0003$ 和 $0.043<0.01$ ），同时品种 A_1 和 A_3 之间也有极显著性差异（ $P=0.0010<0.01$ ）。

从表 8-54 可见，不同品种之间， B_4 的平均产量最高，它与 B_1 、 B_2 和 B_3 之间有显著

性差异 ($P=0.000$ 、 0.000 和 $0.023<0.05$), B_3 与 B_1 、 B_2 之间有显著性差异 ($P=0.023$ 和 $0.008<0.05$), 而 B_2 和 B_1 之间无显著性差异 ($P=0.624>0.05$)。

(8) 交互作用分析

在上述前 4 步的基础上, 单击 Plots 按钮, 打开 Plots 选项卡, 见图 8-11。将施肥变量移入到 Horizontal Axis 下框中, 设定施肥变量为输出图形的横轴, 将品种变量移入到 Separate Lines 下框中, 要求输出不同品种单独的线图。单击 Add 按钮, 则在 Plots 下框中出现所要的设置, 即要求输出不同施肥水平和不同品种间的交互作用图。单击 Continue 按钮, 返回 GLM Univariate 对话框。

单击 OK 按钮运行, 则在输出窗中, 在原有输出的基础上, 增加一张所需的交互作用图, 见图 8-17。图形的纵轴为产量的均值。

从图 8-17 可见, 4 条交互作用线之间彼此交叉, 说明施肥不同水平和不同品种间的交互作用对产量有影响。当品种为 A_2 时、施肥量为 B_4 , 及当品种为 A_3 时、施肥量为 B_3 时, 产量均值达到最大, 这也就是它们现有水平中的最佳搭配方案。

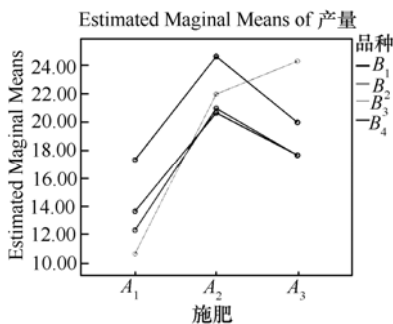


图 8-17 施 N 肥与品种的交互作用图

(9) 结论

由上分析可知, 区组间有显著性差异, 说明土壤的肥力有差异, 施肥因素、品种因素及其它们间的交互作用也有显著性意义。在施 N 肥因素方面, A_2 均值最高, 在品种因素方面, B_4 的平均产量最高, 综合交互作用来看, 采用 A_2B_4 和 A_3B_3 搭配时可获得产量的最大值。

8.3.2 单因变量三因素完全随机试验的方差分析

8.3.2.1 单因变量三因素无重复完全随机试验的方差分析

1. 概念

在三因素试验中, 设 A 因素有 a 个水平, B 因素有 b 个水平, C 因素有 n 个水平, 则这样的试验中共有 abn 个观察值。设 x_{ijk} 为其中的任意一个观察值, 则在该试验中获得的数据模型可参见表 8-41, 只需将表中的区组对应改为 C 因素即可。

由于试验中没有安排重复试验, 因此不能讨论因素间的交互作用。

2. 数据结构

在本模型下获取的数据资料观察值的线性模型可表示为

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}$$

$$(i=1,2,\cdots,a; j=1,2,\cdots,b; k=1,2,\cdots,n)$$

其中, μ 为全部试验观察值的总体均值; α_i 为 A 因素第 i 个水平的效应; β_j 为 B 因素第 j 个水平的效应; γ_k 为 C 因素第 k 个水平的效应; ε_{ijk} 为随机误差, 相互独立且服从 $N(0, \sigma^2)$ 。

3. 三因素方差分析的原假设为

$$H_{01}: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

$$H_{02}: \beta_1 = \beta_2 = \cdots = \beta_b = 0$$

$$H_{03}: \gamma_k = 0, k=1,2,\cdots,n$$

4. 三因素无重复试验方差分析的平方和分解原理

令

$$SS_T = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (x_{ijk} - \bar{x})^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n x_{ijk}^2 - \frac{T^2}{abn}$$

$$SS_A = bn \sum_{i=1}^a (\bar{x}_{i..} - \bar{x})^2 = \frac{1}{bn} \sum_{i=1}^a T_{i..}^2 - \frac{T^2}{abn}$$

$$SS_B = an \sum_{j=1}^b (\bar{x}_{.j.} - \bar{x})^2 = \frac{1}{an} \sum_{j=1}^b T_{.j.}^2 - \frac{T^2}{abn}$$

$$SS_C = ab \sum_{k=1}^n (\bar{x}_{..k} - \bar{x})^2 = \frac{1}{ab} \sum_{k=1}^n T_{..k}^2 - \frac{T^2}{abn}$$

则根据总偏差平方和分解原理, 总有: $SS_T = SS_A + SS_B + SS_C + SS_e$ 的恒等式可知, $SS_e = SS_T - SS_A - SS_B - SS_C$ 。

其中, SS_T 为总偏差平方和, SS_A, SS_B, SS_C 分别为因素 A、因素 B 和因素 C 的主效应平方和, SS_e 为误差平方和, 它是由随机波动引起的。

5. 方差分析表

根据以上的分析, 可以得到如表 8-55 所示的方差分析表。

6. 多重比较

若用 LSD 法进行多重比较, 则当因素 A 各水平之间进行多重比较时,

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{bn}}, \text{ B 因素各水平间进行多重比较时, } S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{an}}, \text{ C 因素各水平间进行}$$

多重比较时, $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{ab}}$, t 分布的自由度为 $df = abn - a - b - n + 2$ 。

表 8-55 三因素试验结果的方差分析用表

方差来源	偏差平方和	自由度	均方	F
因素 A	SS_A	$a - 1$	$S_A^2 = \frac{SS_A}{a - 1}$	$F_1 = \frac{S_A^2}{S_e^2}$
因素 B	SS_B	$b - 1$	$S_B^2 = \frac{SS_B}{b - 1}$	$F_2 = \frac{S_B^2}{S_e^2}$
因素 C	SS_C	$n - 1$	$S_C^2 = \frac{SS_C}{n - 1}$	$F_3 = \frac{S_C^2}{S_e^2}$
误差	SS_e	$abn - a - b - n + 2$	$S_e^2 = \frac{SS_e}{abn - a - b - n + 2}$	
总	SS_T	$abn - 1$		

7. 实例分析

例 8.10 在生产某种特殊物质的过程中, A 表示操作工人 (3 名), B 表示试验所用的催化剂 (3 种), C 表示生产冷却过程的洗涤时间 (15 分钟和 20 分钟), 在每种因素的组下生产产品量见表 8-56, 数据存放在 data08-12.sav 中, 试对其进行方差分析。

表 8-56 三因素试验结果

操作工人	洗涤时间					
	15 分钟			20 分钟		
	催化剂			催化剂		
	1	2	3	1	2	3
1	11.3	10.5	12.0	11.5	10.3	11.0
2	11.5	10.5	10.2	10.9	9.9	11.5
3	14.5	11.5	11.5	12.7	10.9	12.2

【题析】 由于在本例中影响生产产品量的因素有工人因素、催化剂因素和洗涤时间因素三个, 无重复安排, 因此它是一个三因素无重复的试验。

具体解题步骤如下:

(1) 在数据编辑窗口中, 打开数据文件 data08-12.sav。

(2) 按 Analyze → General Linear Model → Univariate 顺序, 打开 Univariate 主对话框, 见图 8-1。

(3) 在左侧因素和协变量框中, 选择 *产品量* 变量, 单击向右箭头, 将其送入 **Dependent Variable** 框; 用同样方法分别将 *操作工人*、*催化剂* 和 *洗涤时间* 三个因素变量送入 **Fixed Factor(s)** 框中。

(4) 单击 **Model** 按钮, 打开 **Model** 对话框, 见图 8-2。在 **Specify Model** 选项中, 选择 **Custom** 选项, 要求对模型进行自定义。

在 **Build Term(s)** 的 **Type** 下拉式选项卡中, 选择 **Main effects** 项, 要求在模型中只考虑因素的主效应。

在左侧因素和协变量框中, 分别选择 *操作工人*、*催化剂* 和 *洗涤时间* 变量将其移入到 **Model** 框中。关闭 **include intercept in model** 选项, 其他保持系统默认选项, 单击 **Continue** 按钮, 返回 **Univariate** 主对话框。

(5) 单击 **OK** 按钮运行, 则在输出窗中, 得到两张表, 其中第二张表为所需的两因素之间的方差分析表, 见表 8-57。

(6) 结果与讨论

从表 8-57 可见, 在操作工人因素、催化剂因素对生产产品量无影响的原假设下, 出现目前统计量的值或者更极端值的概率为 0.008 和 0.010, 故拒绝原假设, 而认为它们对生产产品量的影响均有极显著性意义, 而在洗涤时间因素对生产产品量无影响的原假设下, 出现目前统计量的值或者更极端值的概率

为 0.388, 大于 0.05, 故认为洗涤时间因素对产品量的影响无显著性意义。

(7) 多重比较

单击 **Post Hoc** 按钮, 展开 **Post Hoc** 多重比较对话框, 见图 8-6。在 **Factor(s)** 下框中, 将 *操作工人* 和 *催化剂* 变量移入到 **Post Hoc Tests for** 下框中, 选择等方差假定前提下的 **LSD** 选项, 用 **LSD** 检验法进行多重比较。单击 **Continue** 按钮, 返回 **GLM Univariate** 对话框。

单击 **OK** 按钮运行, 则在输出窗中, 得到四张表, 其中第三张表为不同工人生产产品量均值间的多重比较表, 见表 8-58, 第四张表为不同催化剂对应产品量的均值间的多重比较表, 见表 8-57。

从表 8-58 可见, 3 号工人生产产品量的均值最高, 它与 1 号工人和 2 号工人生产的产品量均值之间均有显著性差异 ($P=0.015$ 和 $P=0.003$, 均小于 0.05), 而 1 号工人和 2 号工人生产的产品量均值之间无显著性差异 ($P=0.393>0.05$)。

从表 8-59 可见, 不同催化剂水平之间, 催化剂 1 水平对应的产品量的均值最高, 它与催化剂 2 水平之间有极显著性差异 ($P=0.003<0.01$), 其他水平之间均无显著性差异。

表 8-57 三因素方差分析结果

Tests of Between-Subjects Effects					
Dependent Variable: 产品量					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	2334.963 ^a	6	389.161	831.441	.000
操作工人	7.041	2	3.521	7.522	.008
催化剂	6.471	2	3.236	6.913	.010
洗涤时间	.376	1	.376	.802	.388
Error	5.617	12	.468		
Total	2340.580	18			

a. R Squared = .998 (Adjusted R Squared = .996)

表 8-58 不同工人生产产品量均值间的多重比较

① 操作工人	② 操作工人	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	.3500	.39499	.393	-.5106	1.2106
	3	-1.1167 [*]	.39499	.015	-1.9773	-.2561
2	1	-.3500	.39499	.393	-1.2106	.5106
	3	-1.4667 [*]	.39499	.003	-2.3273	-.6061
3	1	1.1167 [*]	.39499	.015	.2561	1.9773
	2	1.4667 [*]	.39499	.003	.6061	2.3273

Based on observed means.
The error term is Mean Square(Error) = .468.

*. The mean difference is significant at the .05 level.

表 8-59 不同催化剂对应产品量均值间的多重比较

① 催化剂	② 催化剂	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	1.4667 [*]	.39499	.003	.6061	2.3273
	3	.6667	.39499	.117	-.1939	1.5273
2	1	-1.4667 [*]	.39499	.003	-2.3273	-.6061
	3	-.8000	.39499	.066	-1.6606	.0606
3	1	-.6667	.39499	.117	-1.5273	.1939
	2	.8000	.39499	.066	-.0606	1.6606

Based on observed means.
The error term is Mean Square(Error) = .468.

*. The mean difference is significant at the .05 level.

8. 交叉设计及其实例分析

(1) 基本概念

交叉设计常用于医学试验中，它通常可以考察一个试验因素和两个区组因素（个体差异因素、阶段因素）对观察结果的影响。

当试验中涉及到一个具有 2 水平的试验因素，根据专业知识的需要，希望试验因素的两个水平先后作用于同一个被试对象，且希望试验因素的两个水平在条件相近的同一对被试对象中实施，或在两组被试对象中交叉实施，前者称为配对二阶段交叉设计，后者称为成组二阶段交叉设计，它们统称为 1 次交叉设计或 2×2 交叉设计。

当试验中涉及到一个具有 3 水平的试验因素（处理）分三个不同的阶段先后给予同一个被试对象，观察被试对象接受每种处理后的反应时，这种对应的试验设计的方法称为 3×3 交叉设计。显然，为了消除固定顺序所带来的顺序误差对试验结果的影响，安排试验时，处理的 3 个水平的安排顺序应随机排列，这样至少应安排 $3! = 6$ 组试验，即在 6 组试验中各水平安排的顺序分别为 123, 132, 213, 231, 312, 321。

这种试验的一个显著特点是每个被试对象都要经历两个或 3 个阶段的试验，因此，对每个被试对象而言，均有一个“洗脱期”，即在经过前一阶段的试验后都要经过一段时间的效应去除期，再做下一阶段试验，以消除处理的顺序效应。

由于在交叉设计中，所要考虑的因素总共为 3 个，一个试验因素（处理）加两个对试验结果有影响的非试验因素（个体差异因素和阶段因素），因而，这也属于三因素方差分析的模型所要解决的问题。

(2) 实例分析

例 8.11 配对二阶段交叉设计

某麻醉科医师为研究催醒宁对氟哌啶的作用，选用生理盐水（记为 C）作为催醒宁（记为 T）的对照药，用 6 只老鼠作为被试对象，他将 6 只老鼠按某些条件配成 3 对，每只老鼠接受两次试验，每次试验先经腹腔注射氟哌啶 40ml/kg 和硫酸阿托品 0.05ml/kg；在大鼠入睡后 15 分钟，再经腹腔注射一种药物（催醒宁 T 或生理盐水 C）使其清醒。每对老鼠两次注射催醒药物的先后顺序用随机的方法决定。观察出现蹬和走动

的时间（分），试验结果见表 8-60，数据存放在 data08-13.sav 中，试对其进行方差分析。

表 8-60 试验方案和结果

动物号	1		2		3		4		5		6	
第 1 次试验	T	15	C	31	T	17	C	30	C	25	T	15
第 2 次试验	C	27	T	25	C	28	T	14	T	18	C	26

具体解题步骤如下：

(1) 在数据编辑窗口中，打开数据文件 data08-13.sav。

(2) 按 Analyze → General Linear Model → Univariate 顺序，打开 Univariate 主对话框，见图 8-1。

(3) 在左侧因素和协变量框中，选择 *催醒时间* 变量，单击向右箭头，将其送入 Dependent Variable 框；用同样方法分别将 *处理*、*阶段* 变量送入 Fixed Factor(s) 框中，将 *动物号* 变量送入 Random Factor[s] 框中。

(4) 单击 Model 按钮，打开 Model 对话框，见图 8-2。在 Specify Model 选项中，选择 Custom 选项，要求对模型进行自定义。

在 Build Term(s) 的 Type 下拉式选项卡中，选择 Main effects 项，要求在模型中只考虑因素的主效应。

在左侧因素和协变量框中，分别选择 *处理*、*阶段* 和 *动物号* 变量将其移入到 Model 框中。关闭 include intercept in model 选项，其他保持系统默认选项，单击 Continue 按钮，返回 Univariate 主对话框。

(5) 单击 OK 按钮运行，则在输出窗中，得到三张表，其中第二张表为所需的两因素之间的方差分析表，见表 8-61。

(6) 结果与讨论

从表 8-61 可见，在阶段和动物个体差异因素对催醒时间无影响的原假设下，出现目前统计量的值或者更极端值的概率为 0.632 和 0.421，都大于 0.05，故不拒绝原假设，而认为阶段和动物个体差异因素对催醒时间无显著性意义，而在处理因素对催醒时间无影响的原假设下，出现目前统计量的值或者更极端值的概率为 0.048，小于 0.05，故认为处理因素对催醒时间有显著性意义。

由于处理只有两个水平，因此无须进行多重比较，就可以确定这两个水平之间在催醒时间均值上有显著性差异。

按 Analyze → Compare Means → Means 顺序，展开 Means 对话框，选择 *催醒时间* 变量，将其移入到 Dependent List 下框中，选定 *处理* 变量并将其移入到 Independent List 下框中，单击 OK 按钮运行，则在输出窗口中得到两个水平对应的催醒时间的均值表，见表 8-62。

表 8-61 方差分析表

Tests of Between-Subjects Effects					
Dependent Variable: 催醒时间					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
处理	Hypothesis 200.083	1	200.083	7.950	.048
	Error 100.667	4	25.167 ^a		
阶段	Hypothesis 6.750	1	6.750	.268	.632
	Error 100.667	4	25.167 ^a		
动物号	Hypothesis 159.417	5	31.883	1.267	.421
	Error 100.667	4	25.167 ^a		

a. MS(Error)

表 8-62 两个处理水平对应的催醒时间均值表

Report			
催醒时间			
处理	Mean	N	Std. Deviation
C	25.5000	6	6.09098
T	17.3333	6	4.03320
Total	21.4167	12	6.51513

从表 8-60 可见，催醒宁 T 的催醒时间显著地低于生理盐水 C 的催醒时间。

(7) 结论

处理对催醒的作用有显著性意义，催醒宁 T 的催醒时间显著地低于生理盐水 C 的催醒时间，而阶段和动物个体差异因素对催醒时间无显著性意义。

例 8.12 成组二阶段交叉设计

某公司在原生产药物 B 的基础上，对其剂型进行改造，研发生产出药物 A，为评价 A 药和 B 药是否有相等的生物利用度，选取 24 名被试对象，并将他们随机均分成两组，第 1 组被试对象在第一与第二试验周期分别接受 A 药与 B 药，第 2 组则反之，在两个试验周期期间设立“洗脱期”，用 AUC（血液浓度-时间曲线下面积）作为评价指标，设计与实测结果见表 8-63。表中数据存放在 data08-14.sav 中，试对其进行方差分析。

表 8-63 A、B 两种剂型药物影响下的血液浓度-时间曲线下面积（ $\mu\text{g/ml}$ ）

组别	被试对象号	AUC		组别	被试对象号	AUC	
		周期 1	周期 2			周期 1	周期 2
AB	1	84.4	70.9	BA	3	66.6	63.6
	2	101.0	92.7		5	117.0	107.3
	4	72.7	99.0		7	47.6	54.6
	6	86.7	79.9		9	73.9	65.2
	8	86.6	70.6		10	111.9	114.4
	12	65.5	46.7		11	59.7	57.6
	15	67.4	54.6		13	80.8	74.9
	16	58.2	89.1		14	85.8	97.4
	18	45.2	52.9		17	61.4	46.1
	19	72.6	66.5		20	77.9	62.3
	21	65.4	63.0		22	56.8	61.5
	23	61.9	51.9		24	74.7	76.0

具体解题步骤如下：

(1) 在数据编辑窗口中, 打开数据文件 data08-14.sav。

(2) 按 Analyze → General Linear Model → Univariate 顺序, 打开 Univariate 主对话框, 见图 8-1。

(3) 在左侧因素和协变量框中, 选择 AUC 变量, 单击向右箭头, 将其送入 Dependent Variable 框; 用同样方法分别将处理、阶段变量送入 Fixed Factor(s) 框中, 将被试对象变量送入 Random Factor[s] 框中。

(4) 单击 Model 按钮, 打开 Model 对话框, 见图 8-2。在 Specify Model 选项中, 选择 Custom 选项, 要求对模型进行自定义。

在 Build Term(s) 的 Type 下拉式选项卡中, 选择 Main effects 项, 要求在模型中只考虑因素的主效应。

在左侧因素和协变量框中, 分别选择处理、阶段和被试对象变量将其移入到 Model 框中。关闭 include intercept in model 选项, 其他保持系统默认选项, 单击 Continue 按钮, 返回 Univariate 主对话框。

(5) 单击 OK 按钮运行, 则在输出窗中, 得到三张表, 其中第二张表为所需的两因素之间的方差分析表, 见表 8-64。

表 8-64 方差分析表

Tests of Between-Subjects Effects

Dependent Variable: AUC		Type III Sum of Squares	df	Mean Square	F	Sig.
处理	Hypothesis	.241	1	.241	.003	.958
	Error	1831.302	22	83.241 ^a		
阶段	Hypothesis	82.688	1	82.688	.993	.330
	Error	1831.302	22	83.241 ^a		
被试对象	Hypothesis	14190.920	23	616.997	7.412	.000
	Error	1831.302	22	83.241 ^a		

a. MS(Error)

(6) 结果与讨论

从表 8-64 可见, 在处理因素、阶段因素对 AUC 无影响的原假设下, 出现目前统计量的值或者更极端值的概率为 0.958 和 0.330, 两者都大于 0.05, 说明处理因素和阶段因素对 AUC 的影响无显著性意义, 而在被试对象因素对 AUC 无影响的原假设下, 出现目前统计量的值或者更极端值的概率为 0.000, 小于 0.05, 故认为被试对象因素对 AUC 的影响有极显著性意义。

(7) 结论

AUC 的变异主要受到被试对象个体差异的影响, 与处理和阶段因素无关。

例 8.13 3×3 交叉设计

某研究人员为研究 3 种药物 A、B、C 对高血压的治疗效果, 特别希望考察 3 种药物先后用于同一个患者身上会产生何种疗效, 从众多的高血压患者中随机地选取 12 名作为

被试对象,然后随机地将他们均分成6个组,每组患者接受药物的顺序依次为ABC、ACB、BAC、BCA、CAB、CBA,观察每次服药后的血压值,设计方案与实测数据见表8-65,表中数据存放在data08-15.sav中,试对其进行方差分析。

表 8-65 A、B、C 3 种药物用于每一个高血压患者的疗效观测结果

患者号	组别	药物与血压值 (kPa)						
		时期	1		2		3	
1	1		A	174	B	146	C	164
2	1		A	145	B	125	C	130
3	2		A	192	C	150	B	160
4	2		A	194	C	208	B	160
5	3		B	184	A	192	C	176
6	3		B	140	A	150	C	150
7	4		B	136	C	132	A	138
8	4		B	145	C	154	A	166
9	5		C	206	A	220	B	210
10	5		C	160	A	180	B	145
11	6		C	190	B	145	A	160
12	6		C	180	B	180	A	208

【题析】本试验设计涉及一个分A、B、C 3水平的试验因素(药物种类)和两个区组因素:时期因素和患者号,因此,它是一个“3×3 交叉设计”。

具体解题步骤如下:

(1) 在数据编辑窗口中,打开数据文件 data08-15.sav。

(2) 按 Analyze → General Linear Model → Univariate 顺序,打开 Univariate 主对话框,见图 8-1。

(3) 在左侧因素和协变量框中,选择血压变量,单击向右箭头,将其送入 Dependent Variable 框;用同样方法分别将药物、时期变量送入 Fixed Factor(s) 框中,将患者号变量送入 Random Factor[s]框中。

(4) 单击 Model 按钮,打开 Model 对话框,见图 8-2。在 Specify Model 选项中,选择 Custom 选项,要求对模型进行自定义。

在 Build Term(s)的 Type 下拉式选项卡中,选择 Main effects 项,要求在模型中只考虑因素的主效应。

在左侧因素和协变量框中,分别选择药物、时期和患者号变量将其移入到 Model 框中。关闭 include intercept in model 选项,其他保持系统默认选项,单击 Continue 按钮,返回 Univariate 主对话框。

(5) 单击 OK 按钮运行,则在输出窗中,得到三张表,其中第二张表为所需的两因

素之间的方差分析表, 见表 8-66。

(6) 结果与讨论

从表 8-66 可见, 在药物和患者间个体差异因素对血压无影响的原假设下, 出现目前统计量的值或者更极端值的概率为 0.000, 都小于 0.05, 故拒绝两者的原假设, 而认为不同的药物和患者的个体差异对血压有显著性影响, 而在时期因素对血压无影响的原假设下, 出现目前统计量的值或者更极端值的概率为 0.517, 大于 0.05, 故认为不同时期对血压无显著性影响。

由于 3 个不同药物间至少有两个药物之间有显著性差异, 因此, 要对不同药物进行多重比较。

(7) 多重比较

单击 Post Hoc 按钮, 展开 Post Hoc 多重比较对话框, 见图 8-6。在 Factor(s) 下框中, 将药物变量移入到 Post Hoc Tests for 下框中, 选择等方差假定前提下的 LSD 选项, 用 LSD 检验法进行多重比较。单击 Continue 按钮, 返回 Univariate 主对话框。

单击 OK 按钮运行, 则在输出窗中, 得到四张表, 其中第四张表为使用不同药物时血压均值间的多重比较表, 见表 8-67。

表 8-66 方差分析表

Tests of Between-Subjects Effects					
Dependent Variable: 血压					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
药物	Hypothesis 2259.934	2	1129.967	7.527	.004
	Error 2852.312	19	150.122 ^a		
时期	Hypothesis 205.116	2	102.558	.683	.517
	Error 2852.312	19	150.122 ^a		
患者号	Hypothesis 18280.773	11	1661.888	11.070	.000
	Error 2852.312	19	150.122 ^a		

a. MS(Error)

表 8-67 不同药物之间的多重比较

Multiple Comparisons						
血压						
(I) 药物	(J) 药物	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
A	B	18.8889 [*]	5.40281	.002	7.5807	30.1971
	C	6.0794	5.23480	.260	-4.8772	17.0359
B	A	-18.8889 [*]	5.40281	.002	-30.1971	-7.5807
	C	-12.8095 [*]	4.82007	.016	-22.8981	-2.7210
C	A	-6.0794	5.23480	.260	-17.0359	4.8772
	B	12.8095 [*]	4.82007	.016	2.7210	22.8981

Based on observed means.

The error term is Mean Square(Error) = 150.122.

*. The mean difference is significant at the 0.05 level.

从表 8-67 可见, 使用药物 B 时血压值最低, 与使用药物 A 和 C 时均有显著性差异 ($P=0.002$ 和 0.016 均小于 0.05), 而药物 A 和 C 间无显著性差异 ($P=0.260>0.05$)。

(8) 结论

药物和患者个体因素对血压有显著性意义, 时期因素对血压无显著性影响。3 种药物中, 药物 B 的降压效果最好。

8.3.2.2 单因变量拉丁方随机试验的方差分析

1. 概念

在拉丁方设计中, 涉及 3 个因素, 一个为试验因素, 另两个为区组因素, 3 个因素之间的交互作用无统计学意义, 且 3 个因素的水平数相同。

由于人们最早用 k 个拉丁字母来排列这种方阵，故称为拉丁方。

拉丁方设计的基本做法：将试验因素的 k 个水平随机地排成 k 行和列的方阵，使试验因素的每个水平在任何一行和任何一列均仅出现一次，在方阵 k 行的左边安排一个具有 k 个水平的区组因素，并在其 k 列的上方安排另一个具有 k 个水平的区组因素。

确保试验因素的每个水平在任何一行和任何一列均仅出现一次的方法很多，这里介绍一种最基本的做法，不失一般性，假设试验因素有 6 个水平，用拉丁字母 A、B、C、D、E、F 表示，在第一列上用随机的方法，由上到下对 6 个拉丁字母进行排列，在第二列上，由下至上将第一列的拉丁字母上移一位，而将原处于第一行的拉丁字母放到最后一行的位置，其余各列按前一列排定的顺序，按第二列的做法进行，直至左后一列，由此得到的基本的拉丁方设计见表 8-68。

表 8-68 对 6 水平的试验因素的一种拉丁方设计

B	D	A	E	C	F
D	A	E	C	F	B
A	E	C	F	B	D
E	C	F	B	D	A
C	F	B	D	A	E
F	B	D	A	E	C

拉丁方设计不仅可以节省试验次数，同时它要比随机区组设计具有更高的精度。

拉丁方设计常用于农田试验设计，此外还广泛地应用于医学和其他领域科学试验的设计中。

在使用拉丁方设计试验时，需要注意的事项：

① 若在每一行上使用一位被试对象，则要求试验因素对观察指标产生的效应可在设定的时间间隔内可以消失，使被试对象能恢复到原先的状态；

② 应有实践证明，试验中涉及的 3 个因素之间的交互作用可以忽略不计；

③ 若试验因素是药物种类，此时，不要轻易使用拉丁方设计，因为药物往往会使用疗效指标发生不可逆转的变化，此时，各行上测得的疗效指标的数值所受到的影响是彼此不同的，很难真正反映不同药物的独立疗效。

2. 数据结构

在拉丁方设计的试验中，设试验因素有 k 个水平，则横向和纵向上的两个区组因素也都分别有 k 个水平，则这样的试验中共有 k^3 个观察值。设 x_{ij} 代表第 i 行、第 j 列交叉的观察值，若以 t 代表处理，则在该试验中获得的数据的线性模型可表示为

$$x_{ij(t)} = \mu + \beta_i + \kappa_j + \tau_{(t)} + \varepsilon_{ij(t)}$$

$$(i=1,2,\cdots,k; j=1,2,\cdots,k; t=1,2,\cdots,k)。$$

其中, μ 为全部试验观察值的总体均值; β_i 为第 i 个横行的效应; k_j 为第 j 个纵列的效应; $\tau_{(t)}$ 表示处理效应; $\varepsilon_{ij(t)}$ 为随机误差, 相互独立且服从 $N(0, \sigma^2)$ 。

3. 拉丁方试验的方差分析的原假设为

$$H_{01}: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_{02}: k_1 = k_2 = \cdots = k_b = 0$$

$$H_{03}: \tau_{(t)} = 0, t=1,2,\cdots,k$$

4. 三因素无重复试验方差分析的平方和分解原理

令

$$SS_T = \sum_{i=1}^k \sum_{j=1}^k (x_{ij} - \bar{x})^2 = \sum_{i=1}^k \sum_{j=1}^k x_{ij}^2 - \frac{T^2}{k^2} \quad (\text{总偏差平方和})$$

$$SS_c = k \sum_{i=1}^k (\bar{x}_{i\cdot} - \bar{x})^2 = \frac{\sum_{i=1}^k T_{i\cdot}^2}{k} - \frac{T^2}{k^2} \quad (\text{横行区组偏差平方和})$$

$$SS_r = k \sum_{j=1}^k (\bar{x}_{\cdot j} - \bar{x})^2 = \frac{1}{k} \sum_{j=1}^k T_{\cdot j}^2 - \frac{T^2}{k^2} \quad (\text{纵行区间偏差平方和})$$

$$SS_t = \sum_{t=1}^k (\bar{x}_{(t)} - \bar{x})^2 = \frac{1}{k} \sum_{t=1}^k T_{(t)}^2 - \frac{T^2}{k^2} \quad (\text{处理的偏差平方和})$$

则根据总偏差平方和分解原理, 总有: $SS_T = SS_c + SS_r + SS_t + SS_e$ 的恒等式可知, 随机误差 $SS_e = SS_T - SS_c - SS_r - SS_t$ 。

5. 期望均方

拉丁方试验得到的数据资料中各项变异的自由度和期望均方的组成, 见表 8-69。

表 8-69 拉丁方试验的自由度和期望均方

变异来源	自由度	固定模型	随机模型
横行间	$k-1$	$\sigma^2 + k\kappa_\beta^2$	$\sigma^2 + k\sigma_\beta^2$
纵行间	$k-1$	$\sigma^2 + k\kappa_\kappa^2$	$\sigma^2 + k\sigma_\kappa^2$
处理间	$k-1$	$\sigma^2 + k\kappa_\tau^2$	$\sigma^2 + k\sigma_\tau^2$
试验误差	$(k-1)(k-2)$	σ^2	σ^2

表中, k 表示试验因素所分的水平数, κ_β^2 表示固定模型中横行因素效应方差, κ_κ^2 表示固定模型中纵行因素效应方差, κ_τ^2 表示固定模型中处理间效应方差, σ_β^2 表示随机模型中横行因素效应方差, σ_κ^2 表示随机模型中纵行因素效应方差, σ_τ^2 表示固定模型中处理间效应方差, σ^2 表示试验误差方差。

6. 方差分析表

根据以上的分析，可以得到如表 8-70 所示的方差分析表。

表 8-70 拉丁方试验结果的方差分析用表

方差来源	偏差平方和	自由度	均方	F
横行区组	SS_c	$k-1$	$S_c^2 = \frac{SS_c}{k-1}$	$F_1 = \frac{S_c^2}{S_e^2}$
纵行区组	SS_r	$k-1$	$S_r^2 = \frac{SS_r}{k-1}$	$F_2 = \frac{S_r^2}{S_e^2}$
处理	SS_t	$k-1$	$S_t^2 = \frac{SS_t}{k-1}$	$F_3 = \frac{S_t^2}{S_e^2}$
误差	SS_e	$(k-1)(k-2)$	$S_e^2 = \frac{SS_e}{(k-1)(k-2)}$	
总	SS_T	k^2-1		

7. 多重比较

若用 LSD 法进行多重比较，则各处理之间进行多重比较时， $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{k}}$ ， t 分布

的自由度为 $df = (k-1)(k-2)$ 。

8. 实例分析

例 8.14 为了研究 5 个不同剂量的甲状腺对豚鼠甲状腺肿的影响，考虑到鼠的种系和体重对观测指标可能有一定的影响，因此，在设计试验时，将这两个重要的非处理因素一并安排。由专业知识可知，这 3 个因素之间的交互作用可忽略不计，设计的格式和试验后收集得到的资料见表 8-71，表中数据存放在 data08-16.sav 中，试对其进行方差分析。

表 8-71 5 个不同剂量的甲状腺提取液对豚鼠甲状腺重的影响

种系	甲状腺提取液的剂量（字母）【甲状腺重（g/200g 体重）】					
	体重	1	2	3	4	5
1		C(65)	E(85)	A(57)	B(49)	D(79)
2		E(82)	B(63)	D(77)	C(70)	A(46)
3		A(73)	D(68)	C(51)	E(76)	B(52)
4		D(92)	C(67)	B(63)	A(41)	E(68)
5		B(81)	A(56)	E(99)	D(75)	C(66)

【题析】 本例是一个以甲状腺不同剂量作为处理因素，以豚鼠种系和体重作为区组

变量的标准拉丁方设计试验。

具体解题步骤如下：

(1) 在数据编辑窗口中，打开数据文件 data08-16.sav。

(2) 按 Analyze → General Linear Model → Univariate 顺序，打开 Univariate 主对话框，见图 8-1。

(3) 在左侧因素和协变量框中，选择 甲状腺重变量，单击向右箭头，将其送入 Dependent Variable 框；用同样方法分别将 种系、体重、甲状腺提取剂量变量送入 Fixed Factor(s) 框中。

(4) 单击 Model 按钮，打开 Model 对话框，见图 8-2。在 Specify Model 选项中，选择 Custom 选项，要求对模型进行自定义。

在 Build Term(s) 的 Type 下拉式选项卡中，选择 Main effects 项，要求在模型中只考虑因素的主效应。

在左侧因素和协变量框中，分别选择 种系、体重、甲状腺提取剂量变量将其移入到 Model 框中。关闭 include intercept in model 选项，其他保持系统默认选项，单击 Continue 按钮，返回 Univariate 主对话框。

(5) 单击 OK 按钮运行，则在输出窗中，得到二张表，其中第二张表为所需的两因素之间的方差分析表，见表 8-72。

表 8-72 方差分析表

Tests of Between-Subjects Effects					
Dependent Variable: 甲状腺重					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	119710.920 ^a	13	9208.532	109.617	.000
种系	375.760	4	93.940	1.118	.393
体重	908.160	4	227.040	2.703	.081
甲状腺提取剂量	2690.960	4	672.740	8.008	.002
Error	1008.080	12	84.007		
Total	120719.000	25			

a. R Squared = .992 (Adjusted R Squared = .993)

表 8-73 不同甲状腺提取剂量之间的多重比较

Multiple Comparisons						
甲状腺重						
(1) 甲状腺提取剂量	(2) 甲状腺提取剂量	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
A	B	-7.0000	5.79678	.250	-19.6301	5.6301
	C	-9.2000	5.79678	.138	-21.8301	3.4301
	D	-23.6000 [*]	5.79678	.002	-36.2301	-10.9699
	E	-27.4000 [*]	5.79678	.000	-40.0301	-14.7699
B	A	7.0000	5.79678	.250	-5.6301	19.6301
	C	-2.2000	5.79678	.711	-14.8301	10.4301
	D	-16.6000 [*]	5.79678	.014	-29.2301	-3.9699
	E	-20.4000 [*]	5.79678	.004	-33.0301	-7.7699
C	A	9.2000	5.79678	.138	-3.4301	21.8301
	B	2.2000	5.79678	.711	-10.4301	14.8301
	D	-14.4000 [*]	5.79678	.029	-27.0301	-1.7699
	E	-18.2000 [*]	5.79678	.009	-30.8301	-5.5699
D	A	23.6000 [*]	5.79678	.002	10.9699	36.2301
	B	16.6000 [*]	5.79678	.014	3.9699	29.2301
	C	14.4000 [*]	5.79678	.029	1.7699	27.0301
	E	-3.8000	5.79678	.524	-16.4301	8.8301
E	A	27.4000 [*]	5.79678	.000	14.7699	40.0301
	B	20.4000 [*]	5.79678	.004	7.7699	33.0301
	C	18.2000 [*]	5.79678	.009	5.5699	30.8301
	D	3.8000	5.79678	.524	-8.8301	16.4301

Based on observed means.
The error term is Mean Square(Error) = 84.007.
*. The mean difference is significant at the .05 level.

(6) 结果与讨论

从表 8-72 可见，在豚鼠种类和体重间因素对其甲状腺肿无影响的原假设下，出现目前统计量的值或者更极端值的概率为 0.393 和 0.081，都大于 0.05，故不拒绝两者的原假设，而认为不同的豚鼠种系和不同的体重对甲状腺重无显著性影响，而在甲状腺提取剂量因素对其甲状腺肿无影响的原假设下，出现目前统计量的值或者更极端值的概率为

0.002, 小于 0.05, 故认为甲状腺提取剂量因素对甲状腺重有显著性影响。

由于 5 个不同甲状腺提取剂量间至少有两个水平之间存在显著性差异, 因此, 要对不同甲状腺提取剂量进行多重比较。

(7) 多重比较

单击 Post Hoc 按钮, 展开 Post Hoc 多重比较对话框, 见图 8-6。在 Factor(s) 下框中, 将 *药物* 变量移入到 Post Hoc Tests for 下框中, 选择等方差假定前提下的 LSD 选项, 用 LSD 检验法进行多重比较。单击 Continue 按钮, 返回 Univariate 主对话框。

按 OK 按钮运行, 则在输出窗中, 得到三张表, 其中第三张表为使用不同甲状腺提取剂量时甲状腺重均值间的多重比较表, 见表 8-73。

从表 8-73 可见, 使用甲状腺提取剂量 E 时甲状腺重最高, 它与甲状腺提取剂量 D 之间无显著性差异 ($P=0.524>0.05$), 它俩与甲状腺提取剂量 A、B、C 之间均有显著性差异, 而甲状腺提取剂量 A、B、C 之间无显著性差异。

(8) 结论

豚鼠种系和体重因素对甲状腺重无显著性影响, 甲状腺提取剂量因素对甲状腺重有显著性影响, 其中甲状腺提取剂量 E、D 对甲状腺重影响最大, 这两者间无显著性差异, 它俩与提取剂量 A、B、C 之间均有显著性差异, 而甲状腺提取剂量 A、B、C 之间无显著性差异。

8.3.2.3 单因变量三因素等重复完全随机试验的方差分析

在同时安排多个因素的试验中, 试验因素之间的交互作用有时是难免的, 很多时候也是人们所期盼的, 为对三个因素之间的交互作用进行研究, 应在试验中至少安排二次重复试验 ($n \geq 2$), 只有这样才能对因素之间的交互作用进行分析。

1. 概念

设在试验中安排有 A、B、C 三个因素, 其水平分别为: a、b、c, 为在试验中考察因素间的交互作用, 每个试验重复 n 次, 则总共有 $abcn$ 个观察值, 记 A 因素的第 i 个水平与 B 因素的第 j 个水平、C 因素的第 k 个水平在重复 l 次时得到的观察值为 x_{ijkl} , 其中, $i=1,2,\dots,a$, $j=1,2,\dots,b$, $k=1,2,\dots,c$, $l=1,2,\dots,n$ 。同有交互作用的双因素试验相类似, x_{ijkl} 有如下的形式

$$x_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl}$$

其中, $\alpha_i, \beta_j, \gamma_k$ 称为因素 A、B、C 的主效应, $(\alpha\beta)_{ij}, (\alpha\gamma)_{ik}, (\beta\gamma)_{jk}$ 称为二因素的交互效应, $(\alpha\beta\gamma)_{ijk}$ 称为三因素的交互效应。 ε_{ijkl} 是服从 $N(0, \sigma^2)$ 分布且是相互独立的随机变量。

2. 三因素等重复试验的方差分析的原假设为

$$H_{01} : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

$$H_{02} : \beta_1 = \beta_2 = \cdots = \beta_b = 0$$

$$H_{03} : \gamma_1 = \gamma_2 = \cdots = \gamma_c = 0$$

$$H_{04} : (\alpha\beta)_{ij} = 0, i = 1, 2, \cdots, a; j = 1, 2, \cdots, b$$

$$H_{05} : (\alpha\gamma)_{ik} = 0, i = 1, 2, \cdots, a; k = 1, 2, \cdots, c$$

$$H_{06} : (\beta\gamma)_{jk} = 0, j = 1, 2, \cdots, b; k = 1, 2, \cdots, c$$

$$H_{07} : (\alpha\beta\gamma)_{ijk} = 0, i = 1, 2, \cdots, a; j = 1, 2, \cdots, b; k = 1, 2, \cdots, c$$

3. 三因素等重复试验方差分析的平方和分解原理

与有交互作用的双因素方差分析类似, 可将总偏差平方和分解成如下的恒等式

$$SS_T = SS_A + SS_B + SS_C + SS_{A \times B} + SS_{A \times C} + SS_{B \times C} + SS_{A \times B \times C} + SS_E$$

其中

$$\begin{aligned} SS_T &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c \sum_{l=1}^n x_{ijkl}^2 - \frac{T^2}{abcn} \\ SS_A &= \frac{\sum_{i=1}^a T_{i\cdots}^2}{bcn} - \frac{T^2}{abcn} \\ SS_B &= \frac{\sum_{j=1}^b T_{\cdot j \cdots}^2}{acn} - \frac{T^2}{abcn} \\ SS_C &= \frac{\sum_{k=1}^c T_{\cdots k \cdot}^2}{abn} - \frac{T^2}{abcn} \\ SS_{A \times B} &= \frac{\sum_{i=1}^a \sum_{j=1}^b T_{ij\cdots}^2}{cn} - \frac{\sum_{i=1}^a T_{i\cdots}^2}{bcn} - \frac{\sum_{j=1}^b T_{\cdot j \cdots}^2}{acn} + \frac{T^2}{abcn} \\ SS_{A \times C} &= \frac{\sum_{i=1}^a \sum_{k=1}^c T_{i\cdots k \cdot}^2}{bn} - \frac{\sum_{i=1}^a T_{i\cdots}^2}{bcn} - \frac{\sum_{k=1}^c T_{\cdots k \cdot}^2}{abn} + \frac{T^2}{abcn} \\ SS_{B \times C} &= \frac{\sum_{j=1}^b \sum_{k=1}^c T_{\cdot j k \cdot}^2}{an} - \frac{\sum_{k=1}^c T_{\cdots k \cdot}^2}{abn} - \frac{\sum_{j=1}^b T_{\cdot j \cdots}^2}{acn} + \frac{T^2}{abcn} \end{aligned}$$

$$SS_{A \times B \times C} = \frac{\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c T_{ijk}^2}{n} - \frac{\sum_{i=1}^a \sum_{j=1}^b T_{ij\bullet}^2}{cn} - \frac{\sum_{i=1}^a \sum_{k=1}^c T_{i\bullet k}^2}{bn} - \frac{\sum_{j=1}^b \sum_{k=1}^c T_{\bullet jk}^2}{an} \\ + \frac{\sum_{k=1}^c T_{\bullet\bullet k}^2}{abn} + \frac{\sum_{j=1}^b T_{\bullet j\bullet}^2}{acn} + \frac{\sum_{i=1}^a T_{i\bullet\bullet}^2}{bcn} - \frac{T^2}{abcn}$$

式中

$$T = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c \sum_{l=1}^n y_{ijkl} \quad T_{i\bullet\bullet} = \sum_{j=1}^b \sum_{k=1}^c \sum_{l=1}^n y_{ijkl} \\ T_{\bullet j\bullet} = \sum_{i=1}^a \sum_{k=1}^c \sum_{l=1}^n y_{ijkl} \quad T_{\bullet\bullet k} = \sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n y_{ijkl} \\ T_{ij\bullet} = \sum_{k=1}^c \sum_{l=1}^n y_{ijkl} \quad T_{i\bullet k} = \sum_{j=1}^b \sum_{l=1}^n y_{ijkl} \\ T_{\bullet jk} = \sum_{i=1}^a \sum_{l=1}^n y_{ijkl} \quad T_{ijk\bullet} = \sum_{l=1}^n y_{ijkl}$$

4. 方差分析表

在固定模型下，三因素等重复试验方差分析表见表 8-74。

表 8-74 三因素等重复试验数据资料的方差分析表（固定模型）

方差来源	平方和	自由度	均方	F
因素 A	SS_A	$a-1$	$S_A^2 = SS_A / (a-1)$	$F_A = \frac{S_A^2}{S_E^2}$
因素 B	SS_B	$b-1$	$S_B^2 = SS_B / (b-1)$	$F_B = \frac{S_B^2}{S_E^2}$
因素 C	SS_C	$c-1$	$S_C^2 = SS_C / (c-1)$	$F_C = \frac{S_C^2}{S_E^2}$
交互作用 $A \times B$	$SS_{A \times B}$	$(a-1)(b-1)$	$S_{A \times B}^2 = SS_{A \times B} / (a-1)(b-1)$	$F_{A \times B} = \frac{S_{A \times B}^2}{S_E^2}$
$A \times C$	$SS_{A \times C}$	$(a-1)(c-1)$	$S_{A \times C}^2 = SS_{A \times C} / (a-1)(c-1)$	$F_{A \times C} = \frac{S_{A \times C}^2}{S_E^2}$
$B \times C$	$SS_{B \times C}$	$(b-1)(c-1)$	$S_{B \times C}^2 = SS_{B \times C} / (b-1)(c-1)$	$F_{B \times C} = \frac{S_{B \times C}^2}{S_E^2}$
$A \times B \times C$	$SS_{A \times B \times C}$	$(a-1)(b-1)(c-1)$	$S_{A \times B \times C}^2 = SS_{A \times B \times C} / (a-1)(b-1)(c-1)$	$F_{A \times B \times C} = \frac{S_{A \times B \times C}^2}{S_E^2}$
误差	SS_E	$abc(n-1)$	$S_E^2 = \frac{SS_E}{abcn}$	
总和	SS_T	$abcn-1$		

5. 实例分析

例 8.15 在生产某种特殊物质的过程中，A 表示操作工人（3 名），B 表示试验所用的催化剂（3 种），C 表示生产冷却过程的洗涤时间（15 分钟和 20 分钟），在每种因素的组合下重复试验 3 次，生产产品的产量见表 8-75，表中数据存放在 data08-17.sav 中，试对其进行方差分析。

表 8-75 三个因素重复试验下生产产品的产量

A	洗 涤 时 间					
	15 分钟			20 分钟		
	B			B		
	1	2	3	1	2	3
1	10.7	10.3	11.2	10.9	10.5	12.2
	10.8	10.2	11.6	12.1	11.1	11.8
	11.3	10.5	12.0	11.5	10.3	11.0
2	11.4	10.2	12.0	9.8	12.6	10.8
	11.8	10.9	10.5	11.3	7.5	10.2
	11.5	10.5	10.2	10.9	9.9	11.5
3	13.6	12.0	11.1	10.7	10.2	11.9
	14.1	11.6	11.0	11.7	7.5	10.2
	14.5	11.5	11.5	12.7	10.9	12.2

【题析】 本例中共考察 3 个因素：工人因素、催化剂因素和洗涤时间因素对生产产品产量的影响，并安排 3 次重复试验，因而可用有交互作用的三因素方差分析的方法来处理。

具体解题步骤如下：

（1）在数据编辑窗口中，打开数据文件 data08-17.sav。

（2）按 Analyze → General Linear Model → Univariate 顺序，打开 Univariate 主对话框，见图 8-1。

（3）在左侧因素和协变量框中，选择产量变量，单击向右箭头，将其送入 Dependent Variable 框；用同样方法分别将操作工人、催化剂、洗涤时间变量送入 Fixed Factor (s) 框中。

（4）由于本例要考虑 3 个因素所有的交互作用，因此，对模型不需另作设计，直接采用 Model 对话框中的全部默认选项即可，故直接单击 OK 按钮运行，则在输出窗中，得到两张表，其中第二张表为所需的两因素之间的方差分析表，见表 8-76。

（5）结果与讨论

从表 8-76 可见，在操作工人因素、催化剂因素对产品产量无影响的原假设下，出现目前统计量的值或者更极端值的概率为 0.000 和 0.001，均小于 0.01，故操作工人因素、催化剂因素对产品产量有极显著性影响，而在洗涤时间因素各水平间及其他因素各水平组间对产品产量无影响的原假设下，出现目前统计量的值或者更极端值的概率都大于 0.05，故催化剂因素及各个因素间的交互作用对产品产量无显著性影响。

表 8-76 方差分析表

表 8-77 3 个工人之间的多重比较

Tests of Between-Subjects Effects					
Dependent Variable: 产量					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	41.683 ^a	17	2.451	4.078	.000
Intercept	6811.894	1	6811.894	1.134E4	.000
操作工人	13.958	2	6.979	11.614	.000
催化剂	10.191	2	5.096	8.480	.001
洗涤时间	1.156	1	1.156	1.923	.174
操作工人 * 催化剂	4.869	4	1.217	2.025	.112
操作工人 * 洗涤时间	2.974	2	1.487	2.474	.098
催化剂 * 洗涤时间	3.700	2	1.850	3.079	.058
操作工人 * 催化剂 * 洗涤时间	4.815	4	1.204	2.003	.115
Error	21.633	36	.601		
Total	6875.190	54			
Corrected Total	63.296	53			

^a. R Squared = .658 (Adjusted R Squared = .497)

Multiple Comparisons						
产量 LSD						
① 操作工人 A	② 操作工人 A	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	.4333	.25840	.102	-.0907	.9574
	3	-.7944 [*]	.25840	.004	-1.3185	-.2704
2	1	-.4333	.25840	.102	-.9574	.0907
	3	-1.2278 [*]	.25840	.000	-1.7518	-.7037
3	1	.7944 [*]	.25840	.004	.2704	1.3185
	2	1.2278 [*]	.25840	.000	.7037	1.7518

Based on observed means.
The error term is Mean Square(Error) = .601.
*. The mean difference is significant at the 0.05 level.

(6) 多重比较

单击 Post Hoc 按钮，展开 Post Hoc 多重比较对话框，见图 8-6。在 Factor(s)下框中，将操作工人、催化剂变量移入到 Post Hoc Tests for 下框中，选择等方差假定前提下的 LSD 选项，用 LSD 检验法进行多重比较。单击 Continue 按钮，返回 Univariate 主对话框。

单击 OK 按钮运行，则在输出窗中，得到四张表，其中第三张表为使用不同操作工人生产产品产量的均值间的多重比较表，见表 8-77，第四张表为使用不同洗涤剂水平时生产产品产量的均值间的多重比较表，见表 8-78。

从表 8-77 可见，3 号工人生产的产品产量显著高于 1 号和 2 号工人 ($P=0.04$ 和 $P=0.000$)，而 1 号和 2 号工人生产的产品产量的均值间无显著性差异 ($P=0.102$)。

从表 8-78 可见，洗涤剂 1 水平与 3 水平生产的产品产量均值间无显著性差异 ($P=0.083$)，它们与 2 水平之间均有显著性差异 ($P=0.000$ 和 $P=0.026$)。

(7) 结论

操作工人因素、催化剂因素对产品产量有极显著性影响，在操作

表 8-78 不同洗涤剂时生产产品产量均值的多重比较

Multiple Comparisons						
产量 LSD						
① 催化剂	② 催化剂	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	1.0611 [*]	.25840	.000	.5371	1.5852
	3	.4611	.25840	.083	-.0629	.9852
2	1	-1.0611 [*]	.25840	.000	-1.5852	-.5371
	3	-.6000 [*]	.25840	.026	-1.1241	-.0759
3	1	-.4611	.25840	.083	-.9852	.0629
	2	.6000 [*]	.25840	.026	.0759	1.1241

Based on observed means.
The error term is Mean Square(Error) = .601.
*. The mean difference is significant at the 0.05 level.

工人因素方面, 3 号工人生产的产品产量显著高于 1 号和 2 号工人, 而 1 号和 2 号工人生产的产品产量的均值间无显著性差异, 在催化剂因素方面, 洗涤剂 1 水平与 3 水平生产的产品产量均值间无显著性差异, 均显著地高于使用洗涤剂 3 水平所生产的产品产量。

8.4 单因变量协方差分析

8.4.1 单因变量协方差分析基本概述

1. 定义

R. A. Fisher 在其名著《Statistical Methods for Research Workers》中首先提出了协方差分析的应用。他考虑了协方差模型中的一个较简单的例子, 就是在单向分类模型中添加一个因子的回归项, 使模型变成 $y_{ik} = \beta_0 + \beta_i + x_{ik}\gamma + \varepsilon_{ik}$, $i=1, 2, \dots, r, k=k_i=1, \dots, n_i$,

其中试验次数 $n = \sum_{i=1}^r n_i$ 。

他在影响茶树产量的研究中, 假设试验因素 A 共有 r 种处理, 接受了第 i 种处理的第 k 棵茶树的产量用 y_{ik} 表示。可以设想, 当某些处理恰好安排在产茶能力较强的一些树中, 会不可避免地带来试验误差, 故用 X_{ik} 记茶树在接受处理前的产量, 由于茶树的相对产量说明它对年份是稳定的, 因此, X_{ik} 可以用来作为茶树产茶能力的预报因子。添加 X_{ik} 这一因子后, 可以调整各处理下茶树的平均产量, 从而消除因茶树产茶能力的不同而造成的差异, 降低试验误差, 从而能给出较精细的各处理间的对比。

综上所述, 协方差分析是组合方差分析和回归分析的特征而形成的一种统计方法。

它的基本模型为: $y = X_1\beta + X_2\gamma + \varepsilon$ 。从模型中显而易见, 它的主要成分由两部分组成,

其中 $X_1\beta$ 部分相应的自变量为属性因子 (X_1 是 0-1 矩阵), 称它为方差分析部分, 在模型中, 它往往反映试验中能够由人们精心设计和严格控制的部分; 而 $X_2\gamma$ 部分相应的自变量为数量因子 (X_2 的元素可以连续取值), 称 $X_2\gamma$ 为回归分析部分, 它是试验中那部分无法由人们控制和掌握的因素的作用, 这类在协方差模型中被考虑到回归中的自变量称协同变量或干扰变量。因此, 协方差分析是将那些很难控制而又有可能影响试验结果的因素作为协变量, 在排除协变量影响的前提下, 分析控制变量对因变量的影响, 从而更加准确地对控制因素进行评价。由于引进协同变量后, 对模型做统计分析时需要涉及 X 和 Y 的样本协方差的计算, 因而, 这或许是称它为协方差分析的一个原因。

2. 基本方法

设 $y_* = y - X_2\gamma$, 则 $y = X_1\beta + X_2\gamma + \varepsilon$ 可以改写成 $y_* = X_1\beta + \varepsilon$, 当在 y_* 中, 以 γ 的一个适当的估计量 $\hat{\gamma}$ 代替 γ , 可得到一个以 $z = y - X_2\hat{\gamma}$ 为观察值向量的方差分析模型

$$z = X_1\beta + \varepsilon_1$$

而 $\hat{\gamma}$ 可从 $y = X_1\beta + X_2\gamma + \varepsilon$ 中消去方差分析部分后导出的模型

$$P_{X_1^2}y = P_{X_1^2}X_2\gamma + P_{X_1^2}\varepsilon$$

估出, 其正规方程是

$$X_2'P_{X_1^2}X_2\hat{\gamma} = X_2'P_{X_1^2}y$$

由此可以解得 $\hat{\gamma} = (X_2'P_{X_1^2}X_2)^{-1}X_2'P_{X_1^2}y$ 故得

$$z = [I - X_2(X_2'P_{X_1^2}X_2)^{-1}X_2'P_{X_1^2}]y$$

所以, 对模型 $y = X_1\beta + X_2\gamma + \varepsilon$ 中方差分析部分的统计分析, 可以通过纯方差分析的模型: $Z = X_1\beta + \varepsilon_1$ 进行, 由于已经消去 $X_2\gamma$ 项, 可直接应用方差分析中的现成结果。

模型的精度一般用剩余平方和表示。上面模型的剩余平方和为

$$SS_{\varepsilon^*} = y'P_{X_1^2}y - y'P_{\mu}y'$$

当不引进协同变量时, $SS_{\varepsilon} = y'P_{X_1^2}y$, 所以上面式中的 $y'P_{\mu}y$ 项可看作是引进协变量后对模型提高的精度上的收获。

在协方差分析模型中, 同样可讨论回归因子的显著性检验。当 X_2 为 q 列时, 此时, 所要作的原假设为: $H_0: \gamma_{k+1} = \cdots = \gamma_q = 0$ 。记 $X_2 = [X_{21}:X_{22}]$, 其中 X_{21} 为 k 列, 则假设 H_0 成立时的模型为

$$y = X_1\beta + X_{21}\gamma_{(1)} + \varepsilon$$

其中 $\gamma_{(1)} = (\gamma_1, \gamma_2, \cdots, \gamma_k)'$, 其剩余平方和为 $SS_0 = \|P_{(X_1:X_{21})^\perp}y\|^2$, 而模型 $y = X_1\beta + X_2\gamma + \varepsilon$ 的剩余平方和为 $SS_{\varepsilon} = \|P_{(X_1:X_2)^\perp}y\|^2$, 故令 $SS_H = SS_0 - SS_{\varepsilon}$, 可知 SS_H 与 SS_{ε} 独立, 由此可得协变量的检验统计量

$$F = \frac{SS_H}{SS_{\varepsilon}} \cdot \frac{n-r-q}{q-k} \sim F_{q-k, n-r-q}$$

因此, 综合上述分析可知, 在协方差分析中, 可将总的变异平方和分解为控制变量的变异平方和、协变量的变异平方和、控制变量交互项的变异平方和及随机误差平方和。从而可以构造以随机误差均方为分母的 F 检验, 来检验各控制变量、控制变量的交互项和协变量对因变量的效应。

8.4.2 单因变量协方差分析的实例分析

例 8.16 某高血压研究中心开发了三种治疗高血压的方法，为评价这三种疗法的疗效间有无区别，将 18 名高血压患者随机分成 3 组，每组 6 人，分别接受一种疗法进行为期一个月的临床试验，基础数据和试验结果见表 8-79，表中数据存放在 data08-18.sav 中，使对其进行协方差分析。

表 8-79 18 名高血压患者经三种治疗方法治疗前、后的血压情况

患者 编号	疗法	入院治疗前 血压	入院治疗后 血压	患者 编号	疗法	入院治疗前 血压	入院治疗后 血压
1	1	160	120	10	2	150	110
2	1	185	125	11	2	155	125
3	1	155	130	12	2	155	125
4	1	145	110	13	3	160	105
5	1	175	145	14	3	175	150
6	1	175	130	15	3	165	145
7	2	180	135	16	3	155	140
8	2	210	140	17	3	190	155
9	2	220	155	18	3	165	130

【题析】 治疗方法的有效性可以通过自身比较来加以检验。这可以通过第 6 章中介绍过的自身比较的方法来实现。但前提条件是入院前、后的血压值需服从正态分布，在此前提下，可以验证这三种疗法都是有效的 ($P < 0.05$)。而要评价这三种疗法的疗效间有无区别，显然要用到单因素方差分析，由于入院治疗前的血压会影响到治疗后的结果，因此，入院治疗前的血压就是分析中需要考虑的一个重要的协变量。而作为协方差分析需要考虑的前提条件必须得到满足，才可以用协方差分析的方法。

因此，本例的具体解题步骤如下：

(1) 数据资料的正态性检验和方差齐性检验

首先，在数据编辑窗口中，打开数据文件 data08-18.sav。利用第 2 章 2.4 探索分析例 2.48 中介绍的数据资料的正态性和方差齐性检验方法，对入院治疗前和后的血压变量分别进行正态性和方差齐性检验，可得检验结果，见表 8-80 和表 8-81。

从表 8-80 可见，在三个组中，治疗前、后血压变量在两种检验方法下，均不拒绝服从正态分布的原假设 (P 都大于 0.05)。

从表 8-81 可见，在三个组中，治疗前、后血压变量在四种检验方法下，均不拒绝服从方差齐性的原假设 (P 都大于 0.05)。

这说明，对此进行方差分析的前提条件已经得到满足。

表 8-80 对治疗前、后血压变量的正态性检验

Tests of Normality							
	疗法	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
治疗前血压	疗法1	.244	6	.200 [*]	.925	6	.542
	疗法2	.278	6	.161	.847	6	.149
	疗法3	.272	6	.189	.911	6	.443
治疗后血压	疗法1	.262	6	.200 [*]	.862	6	.195
	疗法2	.168	6	.200 [*]	.978	6	.944
	疗法3	.222	6	.200 [*]	.891	6	.325

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

表 8-81 对治疗前、后血压变量的方差齐性检验

Test of Homogeneity of Variance					
		Levene Statistic	df1	df2	Sig.
治疗前血压	Based on Mean	3.007	2	15	.080
	Based on Median	2.188	2	15	.147
	Based on Median and with adjusted df	2.188	2	12.168	.154
	Based on trimmed mean	3.030	2	15	.078
治疗后血压	Based on Mean	.981	2	15	.398
	Based on Median	.638	2	15	.542
	Based on Median and with adjusted df	.638	2	10.971	.547
	Based on trimmed mean	.881	2	15	.435

(2) 协变量和因变量间的相关分析

关于相关分析的具体内容和步骤将在第 10 章相关分析一节中详细进行论述，这里借用第 7 章非参数假设检验的交叉列表检验中，已介绍过的两个等间隔变量的皮尔逊相关系数的计算方法，来作治疗前、后血压变量的相关分析。它同样可用第 10 章相关分析一节中的内容进行分析。

表 8-82 治疗前、后血压变量的相关性检验

Symmetric Measures					
		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Interval by Interval	Pearson's R	.640	.096	3.335	.004 ^c
Ordinal by Ordinal	Spearman Correlation	.697	.124	3.888	.001 ^c
N of Valid Cases		18			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

利用在 7.9.1.9 中介绍的方法，可得到治疗前、后血压变量的相关系数表，见表 8-82。从表中第一行（两个等间隔变量之间的相关）可见，治疗前、后血压变量间的相关系数为 0.640，在两个变量间相关系数为 0 的原假设下，出现目前统计量的值或者更极端值的概率为 0.004，说明治疗前、后的血压间存在极显著的相关。

因而，协变量应与因变量相关的前提条件也得到满足。

(3) 治疗方法与协变量间的交互作用检验

查看治疗方法与协变量间的交互作用有无显著性意义，可以用来检验各组间回归方程的斜率是否相等的前提条件，当治疗方法与协变量间的交互作用无显著性意义时，说明各组间回归方程的斜率是相等的。

按 Analyze → General Linear Model → Univariate 顺序，打开 Univariate 主对话框，见图 8-1。在左侧因素和协变量框中，选择 *治疗后血压* 变量，单击向右箭头，将其送入 Dependent Variable 框；用同样方法分别将 *疗法* 变量送入 Fixed Factor (s) 框中，将 *治疗前血压* 变量送入 Covariate (s) 框中，定义 *治疗前血压* 变量为协变量。选择有交互作用下的自定义模型，对模型进行定义，将 *疗法*、*治疗前血压*、*疗法*治疗前血压* 项输入到模型中。在 Sum of squares 的下拉列表框中选择 Type I 选项，其他采用系统默认选项，单击 OK 按钮运行，则在输出窗中，得到两张表，其中第二张表为所需的有交互作用的方差分析表，见表 8-82。

注解：I 型（分层处理平方和的方法）方差分析的结果和输入变量的先后顺序有关，不同的顺序会得到不同的分析结果。本例中模型重点要检验 *疗法* 单独对 *治疗后血压* 的影

响，因此要将它放在最前面。

表 8-83 交互作用分析结果

Tests of Between-Subjects Effects					
Dependent Variable: 治疗后血压					
Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2038.020 ^a	5	407.604	3.317	.041
Intercept	316012.500	1	316012.500	2.572E3	.000
疗法	258.333	2	129.167	1.051	.380
治疗前血压	1476.270	1	1476.270	12.015	.005
疗法 * 治疗前血压	303.416	2	151.708	1.235	.325
Error	1474.480	12	122.873		
Total	319525.000	18			
Corrected Total	3512.500	17			

a. R Squared = .580 (Adjusted R Squared = .405)

表 8-84 协方差分析结果

Tests of Between-Subjects Effects					
Dependent Variable: 治疗后血压					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1734.604 ^a	3	578.201	4.553	.020
Intercept	1093.210	1	1093.210	8.608	.011
疗法	294.216	2	147.108	1.158	.342
治疗前血压	1476.270	1	1476.270	11.625	.004
Error	1777.896	14	126.993		
Total	319525.000	18			
Corrected Total	3512.500	17			

a. R Squared = .494 (Adjusted R Squared = .385)

从表 8-83 可见，疗效和治疗前血压的交互作用项无显著性意义 ($P=0.325>0.05$)，说明不能拒绝各组间回归方程的斜率是相等的假设。

由上面三步的检验结果表明，本例可以选用协方差分析模型来讨论这三种疗法的疗效间有无区别的问题。

(4) 协方差分析

在(3)中设置模型时，做一些修改，一是选择只有主效应时的自定义模型，二是对模型进行的定义中，去掉疗法*治疗前血压的交互作用项，三是在 Sum of squares 的下拉列表框中选择 Type III 选项，即采用系统默认选项。

单击 OK 按钮运行，则在输出窗中，又增加二张表，最后一张表为所需的协方差分析表，见表 8-84。

从表 8-84 中可见，作为治疗因素而言，三种不同疗法治疗后血压均值间无显著性差异 ($P=0.342>0.05$)，而治疗前血压这个协同变量对治疗后血压的因变量有极显著性意义 ($P=0.004<0.01$)，说明治疗前血压会影响到治疗后血压的变化。

应该注意的是，三种不同疗法治疗后血压均值间无显著性差异，并不表示三种疗法对高血压的治疗没有效果，而只是说明三种疗法是等效的。疗法有无疗效，在本例中可通过自身比较来加以区分。正如题析中所言的，这三种疗法都是有效的，读者可以自行验证。

在多个控制变量和多个协变量情形下，协方差分析的做法同本例已无本质的区别，因此，对它们的分析，在此就不再赘述了。

8.5 重复测量资料的方差分析

8.5.1 重复测量资料方差分析的基本概述

1. 定义

被试对象在接受不同处理后，同一个测试指标（因变量）在不同时间点上进行多次

测量所得的资料,称为重复测量资料。需要注意的是,这里所谓的“重复”同前面提到的重复试验中的“重复”不是同一意义上的重复,确切地讲,这里应称为多个时间点测量资料更为合适。

这种资料的特点是其定量观测指标的数值会随着时间的变化而发生动态变化,并且各个时间点上的数值间并不满足相互独立的假设。因此,不能用前面介绍的方差分析的方法进行直接处理。

2. 协方差结构及重复测量资料方差分析的前提条件

同前面一样,数据矩阵同参数矩阵 \mathbf{B} 有关为 $\mathbf{Y}=\mathbf{XB}+\mathbf{E}$ 。其中, \mathbf{E} 为随机误差矩阵,各行的 \mathbf{E} 是独立的,且第 i 行服从 $N_r(0, w_i^{-1} \Sigma)$ 。

进行重复测量资料的方差分析,除需要满足前面提到的一般方差分析的条件外,它还有两个附加假定:

① $\Sigma = \Sigma_c \otimes \Sigma_1 \otimes \cdots \otimes \Sigma_t$, 其中 Σ_c 是测量的协方差矩阵, \otimes 是克罗内克乘法运算符。

② Huynh 和 Feldt (1970) 条件: 假设 $\sigma_{rs}^{(k)}$ 是 Σ_k ($k=1, \dots, t$) 的第 (r, s) 个元素; 则对于 $r \neq s$, $\sigma_{rr}^{(k)} + \sigma_{ss}^{(k)} - 2\sigma_{rs}^{(k)} = \text{常数}$ 。矩阵满足这个条件导致正交转换的变量具有球形协方差矩阵; 由于这个原因, 假定有时被称为球形假定。有复合对称特性的矩阵 (也就是, 相同的对角元素和相同的非对角元素) 自动满足这个假定。

上面式中的 t 表示组内因素的数量, c 表示测量次数, r 为变量数量, r_k 是第 k 个组内因素的水平数。 $r_k \geq 2, k=1, \dots, t$ 。

所以, 进行重复测量资料的方差分析需要满足协方差阵的球形对称性或复合对称性。

3. 球形检验

常用于球形检验的方法为 Mauchly 球形检验法。Mauchly 球形检验的无效假设为

$$H_0: \mathbf{M}' \Sigma \mathbf{M} = \sigma^2 \mathbf{I}_m$$

对应的被择假设为

$$H_1: \mathbf{M}' \Sigma \mathbf{M} \neq \sigma^2 \mathbf{I}_m$$

其中, $\sigma^2 > 0$ 未被指定, \mathbf{I} 是 $m \times m$ 的单位矩阵, 以及 \mathbf{M} 是以组内效应有关的 $r \times m$ 的正交矩阵。 \mathbf{M} 使用应用于组内因素等间隔多项式对比产生。

Mauchly 的 W 统计量为

$$W = \begin{cases} \frac{|\Xi|}{(\text{trace}(\Xi)/m)^m} & \text{如果 } \text{trace}(\Xi) > 0 \\ \text{系统缺失值} & \text{如果 } \text{trace}(\Xi) < 0 \end{cases}$$

其中, $\Xi = M^1 A M$, $A = (Y - X\hat{B})^T W (Y - X\hat{B})$ 是残差平方和和叉积的 $r \times r$ 矩阵。

因此, 可以得到 χ^2 统计量为

$$c = \begin{cases} -\rho(n-r_X)\log W & \text{如果 } W > 0 \\ \text{系统缺失值} & \end{cases}$$

χ^2 渐近分布的概率为:

当 n 大及在无效假设下, 对于 $n-r_X \geq 1$ 和 $m \geq 2$, 原假设成立的概率可用下式求得:

$$P_r(-\rho(n-r_X)\log W \leq c) = P_r(\chi_f^2 \leq c) + \omega_2(P_r(\chi_{f+4}^2 \leq c) - P_r(\chi_f^2 \leq c)) + O(n^{-3}).$$

其中,

$$f = m(m+1)/2 - 1, \text{ 为自由度; } \rho = 1 - (2m^2 + m + 2)/(6m(n-r_X))$$

$$\omega_2 = (m+2)(m-1)(m-2)(2m^3 + 6m^2 + 3m + 2)/(288m^2(n-r_X)^2\rho^2)$$

所以, 显著性水平可用下面的函数求得

$$1 - \text{CDF.CHISQ}(c, f) - \omega_2(\text{CDF.CHISQ}(c, f+4) - \text{CDF.CHISQ}(c, f)),$$

其中, CDF.CHISQ 是 SPSS 累积 χ^2 分布函数。由于浮点不精确性导致计算值小于等于 0 的情况中, 显著性被设置为 0。

当拒绝球形假设时, 可以使用多元方差分析结果, 也可以进行校正自由度后的 F 检验。对 F 检验统计量中的自由度进行校正有三种方法, 分别是 Greenhouse-Geisser ε 法, Huynh-Feldt ε 法, 以及下限 ε 法。

$$\text{Greenhouse-Geisser } \varepsilon \text{ 统计量为: } \varepsilon_{GG} = \frac{(\text{trace}(S_E))^2}{d \times \text{trace}(S_E^2)}$$

$$\text{Huynh-Feldt } \varepsilon \text{ 统计量为: } \varepsilon_{HF} = \min\left(\frac{nd\varepsilon_{GG} - 2}{d(n-r_X) - d^2\varepsilon_{GG}}, 1\right)$$

$$\text{下限 } \varepsilon \text{ 统计量为: } \varepsilon_{LB} = \frac{1}{d}. \text{ 它是 } \varepsilon \text{ 可能的最小值。}$$

用这三个方法中的“球对称”系数 ε 乘以被试对象内各变异的自由度, 可实现对 F 检验中自由度的校正。

通常当 $\varepsilon < 0.75$ 时, 使用 Greenhouse-Geisser ε 统计量校正, 当 $\varepsilon > 0.75$ 时, 使用 Huynh-Feldt ε 统计量校正。

对于这三种 ε 中任意一种, 校正的显著性水平是 $1 - \text{CDF.F}(F, \varepsilon d_L, \varepsilon d(n-r_X))$, 其中, ε 是三种 ε 中的任意一种。

4. 变异平方和的分解

对于较为简单的两因素不同时间点上得到的“重复”测量资料, 总变异平方和 $SS_{\text{总}}$ 可分解为

$$SS_{\text{总}} = SS_{\text{被试对象间}} + SS_{\text{被试对象内}} = (SS_{\text{处理}} + SS_{\text{个体间误差}}) + (SS_{\text{时间}} + SS_{\text{处理与时间交互}} + SS_{\text{个体内误差}})$$

当因素更多时或有协同变量存在时, 可在此基础上, 参照前面给出的方差分析的模型, 来构造总变异平方分解的平衡式。

5. 在 SPSS 程序中组间效应、组内效应等检验时的基本步骤

(1) 组间效应的检验

检验无组间效应假设的程序使用以下步骤:

① 计算 $M = I_c \otimes M_{1;1} \otimes \cdots \otimes M_{t;1}$, 其中 $M_{k;1}$ 是第 k 个组内因素的比较矩阵 M_k 的第一列。 M 是一个 $r \times c$ 矩阵。

M_k 是第 k 个组内因素的对比矩阵, $k = 1, \dots, t$ 。它是一个维度为 r_k 的方阵。在第一列中的每个元素通常等于 $\frac{1}{r_k}$ 。对多项式对比而言, 每个元素是 $\frac{1}{\sqrt{r_k}}$, 或对使用者指定的对比而言, 每个元素为非零常数。其他列的列和为 0。

② 为包含截距的每个组间效应, 根据指定的平方和类型, 得到 L 矩阵。

③ 计算 $S_H = (L\hat{B}M)'(LGL')(L\hat{B}M)$ 以及 $S_E = M'SM$, 两者都是 $c \times c$ 矩阵。

④ 计算四个多元检验统计量: Wilks λ , Pillai 轨迹, Hotelling-Lawley 轨迹, Roy 最大平方根, 以及相应的显著性水平。还计算单个变量的 F 统计量。

⑤ 重复步骤 2-4, 直到所有组间效应被检验。

(2) 组内效应的多元检验

检验无组内效应假设的程序使用以下步骤:

① 对于第 k 个组内因素, 计算 $M = I_c \otimes A_1 \otimes \cdots \otimes A_t$, 其中, 当在效应中包含第 k 个组内因素时, $A_k = M_{k;2:r_k}$ 是 M_k 的第二至最后一列, 其它情况时, $A_k = M_{k;1}$ 。 M 是一个 $r \times cd$ 矩阵, 其中, d 是在克罗内克乘积 $A_1 \otimes \cdots \otimes A_t$ 中列的数量。一般而言, $d > 1$ 。

② 为每个组内效应, 根据指定的平方和类型, 得到 L 矩阵。

③ 计算 $S_H = (L\hat{B}M)'(LGL')(L\hat{B}M)$ 以及 $S_E = M'SM$, 两者都是 $cd \times cd$ 矩阵。

④ 计算四个多元检验统计量: Wilks λ , Pillai 轨迹, Hotelling-Lawley 轨迹, Roy 最大平方根, 以及相应的显著性水平。还计算单个变量的 F 统计量。

⑤ 为下一个组间效应重复步骤 2—4, 当所有组间效应被使用时, 直接到步骤 6。

⑥ 重复步骤 1—5, 直到所有组内效应被检验。

(3) 组内效应的均值检验

检验无组内效应假设的均值检验程序使用以下步骤:

- ① 采用 M_k ($k=1, \dots, t$) 等间隔多项式对比矩阵。
 - ② 计算 $M = I_c \otimes A_1 \otimes \dots \otimes A_t$, 其中, 当在效应中包含第 k 个组内因素时, $A_k = M_{k;2:r_k}$ 是 M_k 的第二至最后一列, 其他情况时, $A_k = l_{r_k} / \sqrt{r_k}$ 。 M 是一个 $r \times cd$ 矩阵, 其中, d 是在克罗内克乘积 $A_1 \otimes \dots \otimes A_t$ 中列的数量。一般而言, $d > 1$ 。
 - ③ 为每个组间效应, 根据指定的平方和类型, 得到 L 矩阵。
 - ④ 计算 $S_H = (L\hat{B}M)'(LGL')(L\hat{B}M)$ 以及 $S_E = M'SM$, 两者都是 $cd \times cd$ 矩阵。
 - ⑤ 把 S_H 分割成每个维度为 $d \times d$ 的 c_2 块矩阵。第 (k, l) 块矩阵标识为 $S_{H;k,l}$, ($k=1, \dots, c$ 及 $l=1, \dots, c$), 是从 $(k-1)d+1$ 行至 kd 行和从 $(l-1)d+1$ 列到 ld 列的 S_H 的子矩阵。形成 $c \times c$ 矩阵, 用 \bar{S}_H 标识, 其 (k, l) 元素是 $S_{H;k,l}$ 的轨迹。同样地, 获得矩阵 \bar{S}_E 。
 - ⑥ 使用 \bar{S}_H 和 \bar{S}_E 计算四个多元检验统计量: Wilks λ , Pillai 轨迹, Hotelling-Lawley 轨迹, Roy 最大平方根, 以及相应的显著性水平。
- 在计算中为 \bar{S}_H 设定自由度等于 dr_L 和 \bar{S}_E 的自由度等于 $d(n-r_x)$ 。还计算单个变量的 F 统计量及其显著性水平。
- ⑦ 为每个组间效应重复步骤 3—6。当所有组间效应被使用时, 直接到步骤 8。
 - ⑧ 重复步骤 2—7, 直到所有组内效应被检验。

8.5.2 重复测量资料实例分析

例 8.17 为研究减肥新药盐酸西布曲明片和盐酸西布曲明胶囊的减肥效果是否不同, 以及肥胖患者服药后不同时间的体重随时间的变化情况。采用双盲双模拟随机对照试验, 将体重指数 BMI_{f27} 的肥胖患者 40 名随机等分成两组, 一组给以盐酸西布曲明片+盐酸西布曲明胶囊, 另一组给以盐酸西布曲明胶囊+盐酸西布曲明片。所有患者每天坚持服药, 共服药 6 个月, 试验期间禁用任何影响体重的药物, 而且被试对象行为、饮食及运动与服药前的平衡期均保持一致。分别于平衡期 (0 周)、服药后的 8 周、16 周和 24 周测定肥胖患者的体重 (kg), 得到表 8-85 所示的资料, 表中数据存放在 data08-19.sav 中, 使对其进行方差分析。(资料来源: 方积乾《卫生统计学》第 5 版 P161)

具体解题步骤如下:

(1) 数据资料的正态性检验和方差齐性检验

首先, 在数据编辑窗口中, 打开数据文件 data08-19.sav。利用第 2 章 2.4 探索分析例 2.48 中介绍的数据资料的正态性和方差齐性检验方法, 对 4 个不同时间点测量的体重变量分别进行正态性和方差齐性检验, 可得检验结果, 见表 8-86 和表 8-87。

表 8-85 肥胖患者服用减肥药盐酸西布曲明后的体重 (kg) 变化

被试对象	剂型	服药后测定时间 (周)				被试对象	剂型	服药后测定时间 (周)			
		0	8	16	24			0	8	16	24
1	1	84.4	82.2	82.2	83.0	21	2	64.4	61.4	61.8	62.0
2	1	105.0	100.8	97.4	96.6	22	2	91.0	88.4	87.4	89.6
3	1	63.8	62.0	61.6	60.4	23	2	76.0	76.2	72.8	71.6
4	1	86.2	85.5	83.0	81.8	24	2	71.0	72.0	69.8	68.4
5	1	75.6	73.4	74.0	73.0	25	2	69.4	66.6	62.8	60.8
6	1	61.2	60.4	60.8	60.2	26	2	89.9	87.4	92.6	95.5
7	1	67.8	66.0	63.4	63.6	27	2	66.8	63.6	62.6	61.6
8	1	77.2	73.6	72.6	72.0	28	2	63.4	61.2	62.6	62.0
9	1	73.2	72.2	72.2	74.6	29	2	70.0	67.6	69.8	69.4
10	1	65.4	63.6	62.6	60.8	30	2	86.6	84.0	81.4	78.0
11	1	80.0	77.0	72.4	69.4	31	2	90.4	84.4	77.4	71.0
12	1	74.4	77.0	75.2	77.4	32	2	74.8	73.6	72.8	76.6
13	1	82.6	80.4	81.2	79.6	33	2	67.4	64.4	61.0	58.2
14	1	68.6	65.0	63.2	63.4	34	2	84.4	82.2	80.2	75.4
15	1	79.0	77.0	73.8	72.5	35	2	79.0	76.0	76.5	78.5
16	1	69.4	66.8	64.4	60.8	36	2	87.4	83.2	81.2	77.2
17	1	72.6	71.0	68.2	70.2	37	2	68.7	65.8	63.0	66.4
18	1	72.4	72.6	72.8	72.6	38	2	83.0	81.8	78.4	78.4
19	1	75.6	73.4	73.4	72.2	39	2	66.5	64.4	63.4	65.4
20	1	80.0	78.0	76.4	74.8	40	2	64.6	62.6	64.2	62.0

表 8-86 对重复测量的体重变量的正态性检验

Tests of Normality							
		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
服药0周	1	.129	20	.200 [*]	.913	20	.074
	2	.185	20	.070	.889	20	.026
服药8周	1	.130	20	.200 [*]	.923	20	.112
	2	.177	20	.102	.898	20	.038
服药16周	1	.135	20	.200 [*]	.909	20	.062
	2	.197	20	.040	.907	20	.057
服药24周	1	.128	20	.200 [*]	.919	20	.095
	2	.136	20	.200 [*]	.921	20	.105

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

表 8-87 对重复测量的体重变量的方差齐性检验

Test of Homogeneity of Variance					
		Levene Statistic	df1	df2	Sig.
服药0周	Based on Mean	.977	1	38	.329
	Based on Median	.765	1	38	.387
	Based on Median and with adjusted df	.765	1	36.110	.387
	Based on trimmed mean	.965	1	38	.332
服药8周	Based on Mean	1.029	1	38	.317
	Based on Median	1.094	1	38	.302
	Based on Median and with adjusted df	1.094	1	32.319	.303
	Based on trimmed mean	1.069	1	38	.308
服药16周	Based on Mean	.828	1	38	.368
	Based on Median	.824	1	38	.370
	Based on Median and with adjusted df	.824	1	35.991	.370
	Based on trimmed mean	.676	1	38	.416
服药24周	Based on Mean	.370	1	38	.547
	Based on Median	.351	1	38	.557
	Based on Median and with adjusted df	.351	1	37.946	.557
	Based on trimmed mean	.263	1	38	.611

从表 8-86 可见, 使用 $K-S$ 法检验时, 除服药 16 周时、剂型 2 对应的体重数据资料拒绝正态分布的原假设外 ($P=0.04<0.05$), 其他 7 组的体重数据资料均不拒绝正态分布的原假设。

从表 8-87 可见, 四种检验方法检验结果均不拒绝方差齐性的原假设 (P 都大于 0.05)。可以认为方差分析的基本条件在本数据资料中基本能满足。

(2) 球形检验

按 Analyze → General Linear Model → Repeated Measures 顺序, 打开 Repeated Measures 预对话框, 见图 8-18。

在 Within-Subject Factors Name 框中输入“时间”, 在 Number of levels 框中输入 4, 表示重复测量因素为时间、重复次数为 4 次, 即在 4 个不同时间点上进行测试。由于本例中没有嵌套情形, 所以没有别的可定义的了, 可直接单击 Define 按钮, 进入 Repeated Measures 主对话框, 见图 8-19。

在左侧变量框中, 一次性选中服药 0 周、服药 8 周、服药 16 周和服药 24 周变量, 单击向右箭头, 将其送入 Within-Subject Variables 框中, 定义组内 (时间) 变量; 用同样方法分别将剂型变量送入 Between-Subject Factor (s) 框中, 定义组间变量 (处理)。

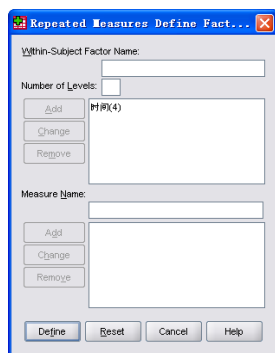


图 8-18 Repeated Measures 预对话框

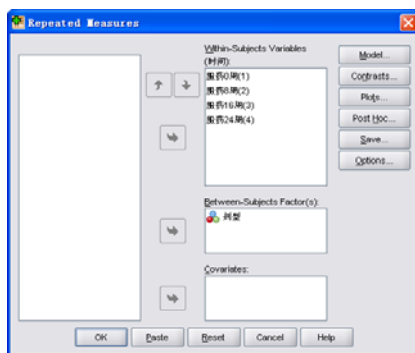


图 8-19 Repeated Measures 主对话框

由于要考虑时间因素和剂型的交互作用, 因此, 可以直接应用模型中的默认选项, 故, 可对系统默认模型设置不做任何修改。直接单击 OK 按钮运行, 则在输出窗口中得到 7 张表的输出结果, 其中第四张表为球形检验结果, 见表 8-88。

从表 8-88 可见, Mauchly 的 W 统计量的值为 0.098, 近似 χ^2 值为 85.133, 自由度为 5, 在球形对称的原假设下, 出现目前统计量的值或者更极端值的概率为 0.000, 故拒绝球形对称的假设, 三种方法的 ε 值依次为 Greenhouse-Geisser ε 统计量的值为: 0.436, Huynh-Feldt ε 统计量的值为: 0.457, 下限 ε 统计量的值为: 0.333。

表 8-88 球形检验

Mauchly's Test of Sphericity ^a						
Measure: MEASURE_1						
Within Subjects Design	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon ^b	
时间	.098	85.133	5	.000	.436	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + 剂型
Within Subjects Design: 时间

表 8-89 多元方差分析

Multivariate Tests ^a						
Effect	Value	F	Hypothesis df	Error df	Sig.	
时间	Pillai's Trace	.715	30.035 ^a	3,000	36,000	.000
	Wilks' Lambda	.285	30.035 ^a	3,000	36,000	.000
	Hotelling's Trace	2.503	30.035 ^a	3,000	36,000	.000
	Roy's Largest Root	2.503	30.035 ^a	3,000	36,000	.000
时间 * 剂型	Pillai's Trace	.043	.541 ^a	3,000	36,000	.657
	Wilks' Lambda	.957	.541 ^a	3,000	36,000	.657
	Hotelling's Trace	.045	.541 ^a	3,000	36,000	.657
	Roy's Largest Root	.045	.541 ^a	3,000	36,000	.657

a. Exact statistic
b. Design: Intercept + 剂型
Within Subjects Design: 时间

球形检验结果表明，在以下分析中，要么使用多元方差分析结果，要么采用校正自由度的 F 检验。

(3) 方差分析

在上面输出窗口的输出结果中，第三张表为多元方差分析结果，见表 8-89，能否用这张表来解释方差分析的结果取决于球形检验的结果，由于球形检验的结果拒绝了球形对称的原假设，因此，可用表 8-89 中的结果来解释本例的方差分析。

从表 8-89 可见，四种检验方法下，时间因素对体重的影响有显著性意义，说明不同时间点测试的体重均值之间至少在两个时间点上显著不同的。而交互作用项对体重的影响不显著。

在上述四种检验方法中，Wiks's Lambda 是常用的检验方法。

表 8-90 组内效应检验表

Tests of Within-Subjects Effects						
Measure: MEASURE_1						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
时间	Sphericity Assumed	384.530	3	128.177	28.213	.000
	Greenhouse-Geisser	384.530	1.308	293.897	28.213	.000
	Huynh-Feldt	384.530	1.372	280.301	28.213	.000
	Lower-bound	384.530	1.000	384.530	28.213	.000
时间 * 剂型	Sphericity Assumed	2.194	3	.731	.161	.922
	Greenhouse-Geisser	2.194	1.308	1.677	.161	.757
	Huynh-Feldt	2.194	1.372	1.599	.161	.768
	Lower-bound	2.194	1.000	2.194	.161	.691
Error(时间)	Sphericity Assumed	517.926	114	4.543		
	Greenhouse-Geisser	517.926	49.719	10.417		
	Huynh-Feldt	517.926	52.130	9.935		
	Lower-bound	517.926	38.000	13.630		

表 8-91 组内效应对照比较检验

Tests of Within-Subjects Contrasts						
Measure: MEASURE_1						
Source	时间	Type III Sum of Squares	df	Mean Square	F	Sig.
时间	Linear	362.882	1	362.882	31.499	.000
	Quadratic	21.609	1	21.609	17.563	.000
	Cubic	.039	1	.039	.045	.834
时间 * 剂型	Linear	1.248	1	1.248	.108	.744
	Quadratic	.576	1	.576	.468	.498
	Cubic	.370	1	.370	.421	.520
Error(时间)	Linear	437.776	38	11.520		
	Quadratic	46.755	38	1.230		
	Cubic	33.395	38	.879		

当然，上述结论也可以从下表中得出。

在上面输出窗口的输出结果中，第五张表为组内效应检验表，见表 8-90。表 8-90 的各因素中的第一行为满足球形假设时的一元方差分析的结果。显然，它在本例中不适用。由于三种方法下的 ε 最大值小于 0.75，所以，使用 Greenhouse-Geisser ε 统计量来校正 F 检验的自由度。因此，表中各因素的第二行的一元方差分析的结果是本例需要重点关注的焦点。表中各因素第二行中的自由度由 $3 \times \varepsilon = 3 \times 0.436 = 1.308$ 得到。同多元方差分析表中得到的结论一样，时间因素对体重的影响有极显著性意义 ($P=0.000<0.01$)，而交互作用项对体重的影响不显著 ($P=0.757>0.05$)。

在上面输出窗口的输出结果中，第六张表为组内效应对照比较检验表，见表 8-91。它是对各次重复测量间变化趋势的分析。除时间因素在线性和二次曲线时(P 均等于 0.000 小于 0.01) 有极显著性意义外，其他情况下，都没有显著性意义。表明 4 个不同时间点上体重变化有线性 and 二次曲线趋势。这一点也可以选用 Plot 对话框做出模型估计的四次测量均值图中得到印证（在 Repeated Measures 主对话框中，见图 8-19，单击 Plots 按钮，在 Profiles Plots 对话框中，将时间变量移入到 Horizontal Axis 框中，单击 Add 按钮，单击 Continue 按钮返回主对话框，单击 Ok 按钮运行，在输出窗口中可得到趋势图），见图 8-20。

在上面输出窗口的输出结果中，第七张表为组间效应检验表，见表 8-92。从表 8-92 中可见，在不同剂型间无差异的原假设下，出现目前统计量的值或者更极端值的概率为 0.897，大于 0.05，因而可以认为不同剂型的减肥效果间没有显著性差异。

表 8-92 组间效应检验表

Tests of Between-Subjects Effects					
Measure: MEASURE_1					
Transformed Variable Average					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	860424.889	1	860424.889	2.485E3	.000
剂型	5.929	1	5.929	.017	.897
Error	13158.052	38	346.265		

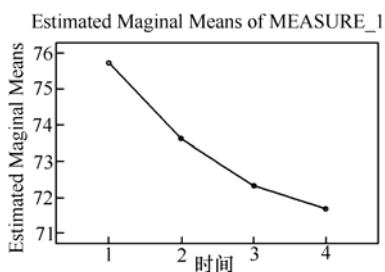


图 8-20 四次测量的趋势图

第9章 正交试验设计与数据分析方法

在实际的科学研究中，影响试验结果的因素往往不是一个，而是多个。例如，从运动训练的角度而言，影响运动成绩的主要因素有运动强度、运动量和运动持续的时间，又如，影响评分性项目的运动员得分的主要因素有运动员水平因素和裁判水平因素等。当总体满足正态分布等一定的条件后，讨论多个因素对试验结果的影响，常用方差分析的方法来进行研究。在用三因素有交互作用的方差分析来研究运动强度、运动量和运动持续的时间对运动成绩的影响时，对影响运动成绩的三个因素各按三个水平进行试验（见表 9-1）。如果进行全面搭配法方案（如图 9-1 所示）安排试验，

表 9-1 因素水平

水平	因素	运动持续 时间（分）	运动量 （米）	运动强度 （秒）
	符号	T	p	m
1		$T_1(100)$	$p_1(4 \times 100)$	$m_1(11.0)$
2		$T_2(150)$	$p_2(5 \times 100)$	$m_2(10.8)$
3		$T_3(200)$	$p_3(6 \times 100)$	$m_3(10.5)$

此方案数据点分布的均匀性极好，因素和水平的搭配十分全面，唯一的缺点是实验次数较多为 $3^3 \times 2 = 54$ 次（指数 3 代表 3 个因素，底数 3 代表每因素有 3 个水平， \times 后面的 2，表示重复一次试验）。因素、水平数越多，则实验次数就越多，例如，做一个 5 因素 3 水平的不重复试验，就需 $3^5 = 243$ 次实验。试验次数越多，就需要更多的人力、物力和财力作保证，而且需要占用更多的时间，这很显然是十分困难的。有时由于所需的时间太长，使试验的条件发生改变，还会导致试验失败；即使试验有了结果，但对运动训练的实际指导意义也可能已经不太大了。因此，需要寻找一种合适的试验设计方法。

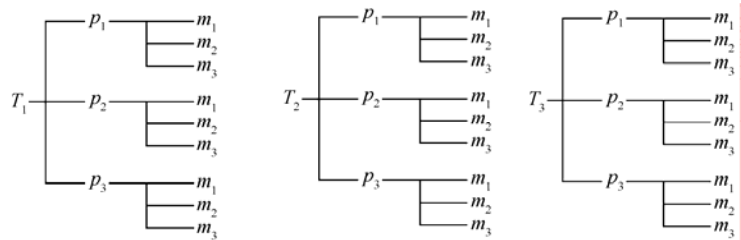


图 9-1 全面搭配法方案

因此，对于如何去做试验，怎样才能做好试验的问题是统计学很关注的一个大问题。这就需要在做具体的试验前，首先需要做好实验设计。

试验设计的一个最重要的原则就是：在做试验前，通过必要的事前考虑，做出合理

周密的事先安排，从而在实际的试验中，通过动用最少的人力、物力、财力及尽可能短的时间，以使用最少的实验次数达到同做大量全面试验等效的结果。

正交试验设计就是在上述的实际需要中逐渐发展成熟并在实际科研工作中被得到广泛运用的一个重要而又有效的统计方法。

9.1 正交试验设计方法的优点和特点

正交实验设计的主要工具是正交表，正交试验设计依托正交表，根据正交性从全面试验中挑选出部分有代表性的点进行试验，这些有代表性的点具备了“均匀分散，齐整可比”的特点，正交试验设计是分析析因设计的主要方法。

9.1.1 正交表

1. 正交表的定义

设 A 是一个 $n \times k$ 矩阵，它的第 j 列元素有数码 $1, 2, \dots, t_j$ $j = 1, 2, \dots, k$ 所构成，如果 A 的任意两列都均匀搭配，则称 A 是一个正交表。

要理解这个概念，最关键的是要理解何谓均匀搭配。简单而言，如果一个矩阵某两列中，同一行中的元素所构成的元素对是一个“完全对”，而且，每对出现的次数相同时，则可说这两列是“均匀搭配”。

假设数组 A 有三个元素为 1, 2, 3，数组 B 有 4 个元素为 1, 2, 3, 4，从数组 A 中任意抽取一个数和从数组 B 中任意抽取一个数，可组成一个元素对，如 (2, 3)，将这两个数组的所有元素对都列出，则可得到“完全对”，如

(1, 1), (1, 2), (1, 3), (1, 4),
 (2, 1), (2, 2), (2, 3), (2, 4),
 (3, 1), (3, 2), (3, 3), (3, 4)。

根据正交表的定义我们知道下面给出的矩阵 A 是一个正交表，因为，该矩阵中任意两列的同行元素都包含 4 个数对 (1, 1), (1, 2), (2, 1), (2, 2)；而 B 不是一个正交表，因为在第 1、第 2 列中的数对为 (1, 1), (2, 2)，各出现两次，而第 1 列和第 4 列的数对为 (1, 1), (2, 2), (1, 2), (2, 1) 为 4 个，各出现 1 次，它不满足任意两列都均匀搭配。

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 1 & 1 & 2 & 2 \\ 2 & 2 & 1 & 1 \end{bmatrix}$$

由此可见，将正交试验选择的水平组合列成表格即可组成正交表，因此，正交表是

一种特制的表格，它最早由日本著名的统计学家田口玄一提出。

2. 正交表的格式

表 9-2 所展示的是一张常用的多因素 2 水平的正交表。当各因素水平相同时，正交表一般用记号 $L_n(m^k)$ 表示，记号中的 L 表示正交表， n 表示需要安排的试验次数， k 表示因素的个数， m 表示各因素的水平数。此时的正交表称为等水平正交表。

对于更一般的情形，正交表可记为 $L_n(m_1 \times m_2 \times \cdots \times m_k)$ ，其中基本的符号的含义同上， $m_1 \times m_2 \times \cdots \times m_k$ 表示试验中最多可安排 k 个因素，其中第 1 个因素的水平数为 m_1 ，第 2 个因素的水平数为 m_2 ，其余类推，并称这种情况下的正交表为混水平正交表。

从表 9-2 的结构和内容来看，在表的各列中，只出现 1 和 2 两个数，因此，它是等水平的正交，它最多可安排二水平的因素有 11 个，需要安排的试验次数为 12，因此，表 9-2 被标识为 $L_{12}(2^{11})$ 。

表 9-2 $L_{12}(2^{11})$ 正交表

列号 试验号											
	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	2	2	2	2	2	2
3	1	1	2	2	2	1	1	1	2	2	2
4	1	2	1	2	2	1	2	2	1	1	2
5	1	2	2	1	2	2	1	2	1	2	1
6	1	2	2	2	1	2	2	1	2	1	1
7	2	1	2	2	1	1	2	2	1	2	1
8	2	1	2	1	2	2	2	1	1	1	2
9	2	1	1	2	2	2	1	2	2	1	1
10	2	2	2	1	1	1	1	2	2	1	2
11	2	2	1	2	1	2	1	1	1	2	2
12	2	2	1	1	2	1	2	1	2	2	1

3. 常见正交表的种类及特点

(1) 等水平正交表

各列水平数均相同的正交表，称等水平正交表，又称单一水平正交表。

常用的 2 水平的正交表有： $L_4(2^3)$ ， $L_8(2^7)$ ， $L_{16}(2^{15})$

常用的 3 水平的正交表有： $L_9(3^4)$ ， $L_{27}(3^{13})$ ， $L_{81}(3^{40})$

常用的 4 水平的正交表有： $L_{16}(4^5)$ ， $L_{64}(4^{21})$ ，...

常用的 5 水平的正交表有： $L_{25}(5^6)$ ， $L_{125}(5^{31})$ ，...

等水平正交表为标准表和非标准表两类，以上列出的都是标准的，标准表有以下几个特点：

① 标准表的结构特点

$$\begin{cases} n_i = m^{1+i} \\ k_i = \frac{n_i - 1}{m - 1} = \frac{m^{1+i} - 1}{m - 1}, \quad i = 1, 2, \dots \end{cases}$$

② 等水平数的正交表，任意两个相邻表具有以下关系

$$\begin{cases} n_{i+1} = nm_i \\ k_{i+1} = n_i + k \end{cases}, \quad i = 1, 2, \dots, I$$

①、②中的 m 是水平数，只要它确定了，则第 i 张标准正交表就随之确定了。因此， m 是构造标准正交表的关键参数。对于任何水平的标准表，由 $i=1$ 来确定最小号的正交表。

③ 用标准正交表可以考察因素间的交互作用。

非标准正交表是为了缩小标准表试验号的间隔而提出来的。常用的非标准表有：

2 水平的正交表： $L_{12}(2^{11})$ ， $L_{20}(2^{19})$ ， $L_{32}(2^{31})$ ，…

3 水平的正交表： $L_{18}(3^7)$

其他水平正交表： $L_{32}(4^9)$ ， $L_{50}(5^{51})$ ，…

非标准表也为等水平表，它同标准正交表的不同之处是非标准正交表不能考察因素之间的交互作用。

(2) 混合水平正交表

各列水平数不相同的正交表，叫混合水平正交表。 $L_8(4^1 \times 2^4)$ 就是一个混合水平正交表名称的写法，从左向右，各符号和数字依次为： L 表示正交表，8 表示需要做试验的次数， 4^1 表示 1 个因素有 4 水平（或 4 水平列的列数为 1）， 2^4 表示 4 个因素有 2 水平（或 2 水平列的列数为 4）。 $L_8(4^1 \times 2^4)$ 常简写为 $L_8(4 \times 2^4)$ 。此混合水平正交表含有 1 个 4 水平列，4 个 2 水平列，共有 $1+4=5$ 列。

常用的混合水平的正交表有： $L_{12}(3 \times 2^4)$ ， $L_{12}(6 \times 2^2)$ ， $L_{16}(4 \times 2^{12})$ ，……，等等，一般在含有正交试验设计内容的统计书后均有这些正交表的附表。

笼统地说，混合水平的正交表不能用来考察交互作用，由标准表通过并列法改造得来的混合型正交表除外，但必须回到原标准表上进行。

4. 正交表的性质

正交表最大的特性是它具有正交性。它表现为：

(1) 在任一系列中各水平都出现，且出现的次数相等。例如，正交表 $L_{12}(2^{11})$ 中，每列不同的数字是 1, 2，它们各出现 6 次。

(2) 任意两列中，将同一行的两个数字看成是有序数对时，每种数对出现的次数是相等的，如在正交表 $L_{12}(2^{11})$ 中，出现的有序数对分别为：(1, 1)，(1, 2)，(2, 1)，(2, 2)，它们各出现 3 次。

以上两点充分地体现了正交表的两大优越性，即“均匀分散性，整齐可比”。通俗地说，每个因素的每个水平与另一个因素各水平各碰一次，这就是正交性。

由正交表的正交性可见：

- (1) 正交表的各列地位均等，可以互换，称为列置换；
- (2) 正交表的各行也可以互换，称为行置换；
- (3) 正交表的同一列的水平数也可以相互置换，称为水平置换。

这三种置换称为正交表的三种初等变换，经过初等变换得到的正交表是原正交表的等价表。实际应用时可根据不同的试验要求，充分利用好这一点。

由于正交表具有正交性的特点，因而使它具有很好的代表性。它体现在两个方面，一方面，任一列的各水平都出现，使得所做的部分试验中包含了所有因素的所有水平；而任意两列的所有水平都出现，可使任意两个因素的所有水平信息及其所有组合信息都能体现，从而可使只做部分试验达到了解全部试验的目的。另一个方面，正交试验的试验点均匀地分布在全面试验点之中，具有很强的代表性，故部分试验寻找到的最优条件与全部试验所找的最条件具有趋势一致性的关系。

也正是由于正交表具有正交性的特点，使得各因素各水平的试验条件相同，这可以最大限度地排除其他因素的干扰，从而可以综合比较该因素不同水平对试验指标的影响，即它具有综合可比性。

正是由于正交表有以上的性质，所以，我们说，用它来安排试验时，各因素的各水平的搭配是均衡的。

9.1.2 正交试验设计方法

1. 正交试验设计方法的定义

用正交表安排多因素试验的方法，称为正交试验设计法。

在正交试验设计中，一个因素在正交表中占用一列。

2. 交互作用

在多因素试验中，由于各因素不仅独立地在起作用，而且各因素还经常联合起来起作用。这也就是说，不仅各因素的水平变化时会对试验指标产生影响，而且各因素的联合搭配对试验指标也有影响。这后一种的影响叫因素的交互作用。因素 A 和因素 B 的交互作用记作 $A \times B$ 。

例如，对四名身体素质和各方面条件基本相似的 16 岁男子少年跳远运动员进行不同方式的力量、速度训练，训练 6 个月后，跳远成绩提高的幅度见表 9-3。

从表中不难看出，单纯进行力量训练时，跳远成绩提高 $22-10=12$ (cm)，单纯进行速度训练时，跳远成绩提高 $28-10=18$ (cm)，而进行力量、速度搭配练习时，除保持原有力量、速度训练单独的效果外，力量、速度训练的搭配练习还可产生联合效果，即跳

交互作用表的查法:

表中有两种数字的表达方式,一种是带括号的,一种不带括号。它们都表示列号。当我们把一个因素放在表中括号中的数字所在的列,把另一个因素放在表头列号的其他数字的所在列时,则这两个列号数字在表中交叉处的数字就是两列的交互作用列所对应的列号。

例如,在两因素二水平的研究中,我们把因素 A 放在第 1 列,因素 B 放在第 2 列,则 $A \times B$ 将安排在 (1) 和列号 2 在表中对应交叉处的 3 列上。又如,我们在 2、5 列上已安排好因素,则这两列的交互作用列安排在第 7 列上。从表中也可看出,第 7 列还是 3、4 列的交互作用列。不同的列的交互列有可能出现在同一列上,这是可以的。安排试验时应使因素列和交互作用列分开,各占一列。

常见的交互作用表还有: $L_8(2^7)$ 、 $L_{27}(3^{13})$ 等。

4. 选择正交表的基本原则

合适的正交表的选择取决于试验中所要考虑的影响试验结果(试验指标)的因素(自变量)的个数、水平数以及是否要考虑因素间的交互作用。所以,一般都是要先确定试验的因素、水平和交互作用,然后选择适用的 L 表。

(1) 选择何种类型的正交表,首先,要考虑因素的水平数。如果各因素全是 2 水平,则选用 $L_m(2^k)$ 表;如果各因素全是 3 水平,则选择 $L_m(3^k)$ 表。如果各因素的水平数不相同,则选择适用的混合水平表。

(2) 其次,要考虑交互作用。两因素的交互作用至少要占一列,究竟要占几列这同产生交互作用的因素的自由度有关。

所谓自由度是指独立数据或变量的个数。

在正交表中,可根据以下两条原则来确定自由度:

① 正交表每列的自由度等于各列的水平数减 1,由于因素和列是等同的,从而每个因素的自由度等于该因素的水平数减 1;

② 两因素交互作用的自由度等于两因素的自由度的乘积,即 $f_{A \times B} = f_A \times f_B$ 。

根据这个原则,可以知道,两水平因素的交互作用列只有一列,因为,它的自由度只有 1,两个 3 水平因素的交互作用列有两列,这是因为,它的自由度为 4,而交互作用也是 3 水平的,所以,它要占两列,由此可得,两个 n 水平因素的交互作用的自由度为 $(n-1)(n-1)$,交互作用也是 n 水平的,因此它要占 $(n-1)$ 列。

(3) 选择的正交表要能容纳所考虑的因素和交互作用。为了对试验结果进行方差分析或回归分析,还必须至少留一个空白列,作为“误差”列,在极差分析中要作为“其他因素”列处理。

(4) 在同水平中取何种试验次数的 L 表,取决于试验精度的要求。如果试验精度要求高,则宜取实验次数多的 L 表。

(5) 要根据研究的成本来决定适合的 L 表的选择。若试验费用很昂贵, 或试验的经费很有限, 或人力和时间都比较紧张, 则应选试验次数少一些的 L 表。

(6) 在按原来考虑的因素、水平和交互作用去选择正交表时, 如无正好适用的正交表可选, 则简便且可行的办法是适当修改原定的水平数。

(7) 在对某些因素间的交互作用的影响是否确实存在没有把握的情况下, 如果条件许可, 则应尽量选用大表, 让影响存在的可能性较大的因素和交互作用各占适当的列。则在用方差分析进行显著性检验时, 就可得出结论。这样既不增加太多试验的工作量, 又不致于漏掉重要的信息。

5. 正交试验设计方法的优点和特点

(1) 正交试验设计法的优点

操作简便, 考虑全面, 数据点分布均匀, 能同全部试验等效, 结论的可靠性较好。

(2) 正交试验设计法的特点

完成试验要求所需的实验次数少; 数据点的分布很均匀; 可用相应的极差分析方法、方差分析方法、回归分析方法等对试验结果进行分析, 引出许多有价值的结论。

9.2 正交试验设计的基本步骤

1. 根据研究目的设计试验因素和试验指标

先根据研究课题来确定研究目的, 再从专业的角度在众多影响研究目的的因素中找出几个主要影响因素, 根据研究精度的要求和课题经费的情况确定因素的水平, 一般在条件允许的前提下, 主要影响因素的水平可以分得多一些, 同时还将确定最能反映试验目的测试指标, 以便通过对试验结果的分析找出主、次影响因素。

2. 做好正交表的表头设计

所谓表头设计, 就是把试验中所要考虑的因素和交互作用, 放在正交表中合适的列上。

(1) 当试验不考虑交互作用时, 表头设计可以是任意的。

(2) 当有交互作用时, 要选用合适的交互作用表, 并严格地按规定做好表头设计。

3. 选择合适的正交表做好试验方案安排

根据正交表的表头设计, 在许多可选的正交表中, 参考 9.2.1 中的要求, 选择符合多个方面要求的试验次数又相对较少的正交表, 并做出试验方案的安排。

4. 选择被试对象

要求被试对象间的条件要基本一致, 被试对象的数量由 9.2.3 中所选择的正交表中试验次数的多少来决定。

5. 进行统计分析并给出统计结论

(1) 极差分析

通过极差分析, 来确定最优方案的搭配。

(2) 方差分析

通过方差分析，来检验各因素和交互作用是否有显著性意义。

9.3 正交试验设计实例

1. 不考虑因素间交互作用的情形

例 9.1 某炼钢厂为了提高铁水温度，需要通过试验来选择最好的方案。经过专家筛选，将主要影响铁水温度的因素集中在焦比、风压和底焦高度上，由于三水平的因素与试验指标趋势图多数为二次曲线，更加有利于从专业角度对试验结果的分析，因而要求每个因素都考虑三个水平，已知这三个因素间无交互作用，试对其进行正交试验方案的设计。

(1) 研究目的

寻找影响铁水温度的最佳条件搭配。

(2) 确定试验指标和试验因素

铁水温度作为试验需要考察的指标，各种搭配的试验方案中对应铁水温度最高者为最试验佳搭配方案。

现分别用字母 A、B、C 表示焦比、风压和底焦高度三个因素，根据专家经验和专业知识所确定的各因素的水平值，见表 9-5。

表 9-5 影响铁水温度因素的水平表

因素 水平	焦比 (A)	风压 (B) (毫米水银柱)	底焦高度 (C) (米)
1	1 : 16	170	1.2
2	1 : 18	230	1.5
3	1 : 14	200	1.3

(3) 正交表表头设计

由于本例已知各因素间无交互作用，因此，根据每个因素要占据一列和至少要列出一个空列作为“误差”列，及尽可能用最少数次试验的设计要求，本例可选择 $L_9(3^4)$ 正交表。

这也可以通过 9.1.1 3 中标准正交表的计算公式计算得到 4 因素 3 水平所需要的最小试验次数的正交表。

由于水平数 $m=3$ ，因素（含误差列） $k=4$ ，所以根据 $k = \frac{n-1}{m-1}$ 可计算得到最少试验次数 $n = k(m-1) + 1 = 4(3-1) + 1 = 9$ ，而 $9 = 3^{1+1}$ ，满足其第一个公式 $n_i = m^{1+i}$ ，故本例可选择 $L_9(3^4)$ 正交表。

表 9-6 列出的表头设计是其中的一种。事实上, 在本例的情况下, A、B、C 三个因素可放在 4 列中的任意 3 列上均可, 空出的 1 列可用它来计算误差项。

表 9-6 例 9.1 的表头设计

因素	A	B	C	
列号	1	2	3	4

(4) 编制试验方案

满足设计基本要求的最小正交表可从相关统计书籍中获取, 当不需要考虑因素间的交互作用时, 也可从 SPSS 的 Data 菜单 Orthogonal Design 过程的 Generate 程序中生成。

在上一步表头设计的基础上, 将所选正交表的各个水平数字换成对应的具体水平值, 则得到相应的试验方案, 它是具体试验的依据。例 9.1 用表 9-6 所做的试验方案见表 9-7。

表 9-7 影响铁水温度 3 因素 3 水平的正交试验方案

因素	A	B	C	误差	试验结果
列号 试验号	1	2	3	4	
1	1 (1:16)	1 (170)	1 (1.2)	1	
2	1	2 (230)	2 (1.5)	2	
3	1	3 (200)	3 (1.3)	3	
4	2 (1:18)	1	2	3	
5	2	2	3	1	
6	2	3	1	2	
7	3 (1:14)	1	3	2	
8	3	2	1	3	
9	3	3	2	1	

本试验中的被试对象为相同性质、相同体积的铁水, 随机地安排在 9 个试验号上。尽可能同时进行试验。如, 第一号试验为: $A_1B_1C_1$, 即焦比 1:16, 分压 170 毫米水银柱, 底焦高度 1.2 米。

至此, 我们已安排好了试验设计方案。接下来就可以严格按试验设计方案进行试验, 获取试验结果。

由于本例不需要考虑交互作用, 所以, 此时的正交表和试验方案也可直接在 SPSS 中生成。

在 SPSS16.0 中, 生成本例四因素 (含一列空白列存放误差项) 三水平正交试验设计的操作步骤如下:

1. 在数据编辑窗口, 新建一个数据文件。
2. 按 Data→Orthogonal Design→Generate 顺序单击菜单项, 展开正交设计对话框, 见图 9-2。

在 Factor Name 框中输入 A，单击 Add 按钮，单击 A (?)，单击 Define Values 按钮，弹出 Define Values 对话框，见图 9-3。在 Auto-Fill 下的 From 1 to 框中输入水平数 3，单击 Fill 按钮，在右侧的 Value 下框中自动出现 1、2、3，单击 Continue 按钮，返回正交设计对话框，完成对 A 的定义，依次逐个输入因素变量名，单击 Define Values 按钮在相应的对话框中定义因素水平值，直至 D 为止。

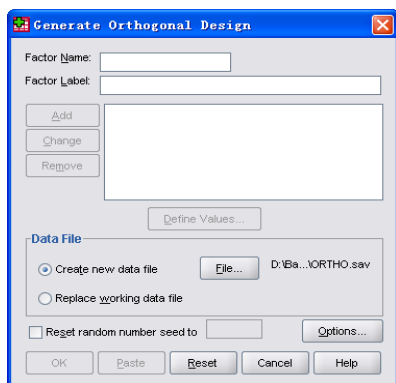


图 9-2 正交设计对话框

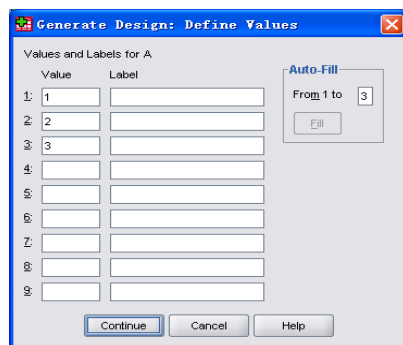


图 9-3 Define Values 对话框

在 Data File 栏内选择 Replace Working data file，设置将设计结果显示在当前工作数据文件中，即当前的数据窗中。

选择 Reset random number seed to，设置随机数种子，随便填写一个正整数：如 2345。随机产生正交设计方案。

单击 OK 按钮提交执行。则在当前工作数据窗中生成正交设计结果。

3. 按 Data→Sort cases 顺序，展开 Sort cases 对话框。

在左边变量名源框中，依次选择 A、B、C，将其移入到 Sort by:框中，在 Sort order 下保持系统默认选择 Ascending（升序），单击 OK 按钮提交执行。则将当前数据文件中生成的正交表进行从小到大的有序排序。见图 9-4。将其存放在 data09-01.sav 的数据文件中。

	A	B	C	D	STATUS_	CARD_
1	1.00	1.00	1.00	1.00	0	9
2	1.00	2.00	2.00	3.00	0	7
3	1.00	3.00	3.00	2.00	0	6
4	2.00	1.00	3.00	3.00	0	1
5	2.00	2.00	1.00	2.00	0	2
6	2.00	3.00	2.00	1.00	0	4
7	3.00	1.00	2.00	2.00	0	5
8	3.00	2.00	3.00	1.00	0	8
9	3.00	3.00	1.00	3.00	0	3

图 9-4 生成的正交设计表

比较图 9-4 和表 9-7 中的各列的水平值，细心的读者一定会发现，两个 4 因素 3 水平的正交表，并不完全一样。

事实上, 将图 9-4 中的第 4 列和第 3 列对调, 再将调整后的第 4 列中水平数 2 和 3 对换, 则可得到和表 9-7 中一样的正交表。

表 9-7 和图 9-4 中的都是正交表, 其中一个是另一个的初等转换的结果, 因此, 它们都可安排正交试验, 所不同的是由它们安排的正交试验的方案有可能是完全一样的。

你或许还会发现, 每次重复上述步骤的运行, 都可能得到不完全相同的正交表, 这是因为程序每次产生的正交表都是随机的。

所以, 如果用统计书中的正交表安排的试验, 一般不能将试验结果直接填在由 SPSS 生成的正交表后的 STATUS_列中, 除非调整表头设计后, 能使各因素对应的水平值都相同, 否则要出现低级错误。正确的做法是将书上正交表中各列水平值, 原封不动地复制到 SPSS 的数据文件中, 是肯定不会有问题的。

尽管, 这两者之间有些不同, 但无论用哪种方法进行正交试验的方案设计都是可行的。

为便于比较, 本例在 SPSS 生成的正交表中, 将因素 C 调整在 D 列, 这样可得到一样的正交试验方案, 并在后续的极差分析和方差分析中也能得到一样的分析结果。

4. 在 SPSS16.0 中生成试验方案

在 SPSS16.0 中, 生成本例 3 因素 3 水平正交试验设计的试验方案的操作步骤如下:

(1) 在数据编辑窗口中, 打开 data09-01.sav。

根据表 9-5 中, A、B、C 三个因素的水平值, 对数据文件中的变量 A、B、D 的值标签分别进行定义, 并将 CAED_变量, 修改为试验号, 在该变量中, 由上到下依次输入 1、2、3、…、9, 将定义好的结果存放在 data09-01a.sav 中。

(2) 按 Data→Orthogonal Design→Display 顺序单击菜单项, 展开 Display Design 对话框, 见图 9-5。

在左边变量名源框中, 选择代表 3 个因素的变量名 A、B、D 试验号, 将其移入到 Factors 框中, 在 Format 选项中选择 Listing for experimenter, 要求输出试验方案。

(3) 单击 OK 按钮, 在输出窗中生成试验方案, 见图 9-6。

它同表 9-7 是一致的。

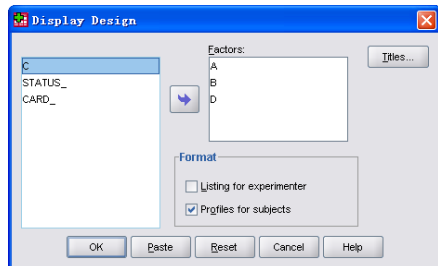


图 9-5 Display Design 对话框

Card List					
	Card	A	B	D	试验号
1		焦比1:16	风压170	底焦高度1.2	1
2		焦比1:16	风压230	底焦高度1.3	2
3		焦比1:16	风压200	底焦高度1.5	3
4		焦比1:18	风压170	底焦高度1.3	4
5		焦比1:18	风压230	底焦高度1.5	5
6		焦比1:18	风压200	底焦高度1.2	6
7		焦比1:14	风压170	底焦高度1.5	7
8		焦比1:14	风压230	底焦高度1.2	8
9		焦比1:14	风压200	底焦高度1.3	9

图 9-6 试验方案

5. 考虑因素间交互作用的情形

从上面的例子中, 我们看到, 在不考虑交互作用时, 正交表可以在 SPSS 中直接生成, 而且只需生成试验因素个数+1 的列数即可。试验次数也会随之而确定。

那么在考虑因素间的交互作用时, 是否也能仿此进行呢? 答案显然不行。

在考虑交互作用时, 一定要按交互作用表做好表头设计, 应采用统计书中提供的与交互作用表相对应的正交表来安排试验, 如果直接采用 SPSS 生成正交表, 则由于书中正交表已经过初等变换, 两种方式下得到的正交表并不完全相同, 因而有可能使得相同试验方案的试验结果, 会得到不同的交互作用的统计分析结果。

当然对在 SPSS 中生成的正交表经初等转换后, 也能变成同统计书中完全一样的正交表, 但这一步很麻烦。为说明这一点, 本例将按这种方式来生成正交表。以此说明它并非最佳方案, 建议对最好不用直接在 SPSS16.0 中生成正交表, 而将正规书中提供的正交表直接作为变量输入到 SPSS16.0 的数据文件中, 或经常用的话, 可在 EXCEL 中先存放好各种正交表, 需要时直接复制粘贴到 SPSS16.0 的数据文件中。

建立数据文件时, 如果只做方差分析, 只需将各试验因素所在的水平值录入到数据文件即可, 但如果要做极差分析, 则需将交互作用列的水平值一并输入到数据文件。空白列不用录入。

例 9.2 为研究不同训练手段及其一级交互作用对男子 800 米自由泳运动员的运动成绩的影响, 试对其进行多因素正交试验方案的设计。

(1) 研究目的

寻找有效提高男子 800 米自由泳成绩的最佳训练手段。

(2) 确定试验指标和试验因素

① 试验因素

根据对专家的调查, 在众多的训练手段中, 确定了包括划臂练习在内的 4 个因素、每个因素分两个水平, 因素的名称和对应水平下的强度的划分见表 9-8。

表 9-8 影响 800 米自游泳成绩的因素和水平值表

因素 水平	划臂练习 A	打腿练习 B	手腿配合游 C	力量练习 D
水平 1	0.5km	1km	4km	10 组
水平 2	0.8km	1.5km	6km	20 组

② 试验指标

800 米自游泳成绩的提高值。

(3) 考虑交互作用

根据经验, 需要考虑因素 A、B、C、D 之间的一级交互作用, 即 $A \times B$ 、 $A \times C$ 、 $A \times D$ 、 $B \times C$ 、 $B \times D$ 、 $C \times D$, 共 6 个。

(4) 选择合适的交互作用表

由于我们需要考虑 4 个因素和 6 个交互作用, 及留出一个随机误差的空白列, 因此, 可以选择 $L_{16}(2^{15})$ 交互作用表 (表 9-4) 来安排试验的表头设计。

前面已经提到过, 考虑交互作用应选用标准正交表。

用样, 这也可以通过 9.1.1 3 中标准正交表的计算公式, 来计算得到 4 因素 3 水平所需要的最小试验次数的标准正交表。

由于水平数 $m=2$, 因素 (不含误差列) 有 4 个, 加上 6 个交互作用项和 1 个空白列, 最少表格要有 11 列, 即 $k=11$, 所以根据 $k = \frac{n-1}{m-1}$ 可计算得到最少试验次数 $n = k(m-1) + 1 = 11(2-1) + 1 = 12$, 而 $12 = 2^{i+1}$, i 必须取正整数, 可得 $i=3$, 因此, $n_i = m^{i+1} = 2^{3+1} = 16$, $k = \frac{16-1}{2-1} = 15$, 故本例可选择 $L_{16}(2^{15})$ 的标准正交表。

(5) 表头设计

根据表 9-4, 首先将因素 A 放在第 1 列, 把因素 B 放在第 2 列, 因第 3 列为 $A \times B$ 交互作用列, 故将因素 C 放在第 4 列, 同理, 将因素 D 放在第 8 列, 由表 9-4 可知, 第 1 列和第 8 列的交互作用列在第 9 列, 因此, $A \times D$ 在第 9 列, 同理, $B \times D$ 在第 10 列, $C \times D$ 在第 12 列, 故 4 个因素和 6 个交互作用的表头设计见表 6。

表 9-9 4 因素之间有交互作用的表头设计

列号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
表头设计	A	B	$A \times B$	C	$A \times C$	$B \times C$		D	$A \times D$	$B \times D$		$C \times D$			

(6) 在 SPSS16.0 中生成试验方案

在 SPSS16.0 中, 生成 15 因素二水平正交试验设计表的操作步骤如下:

① 生成基本的标准正交表

在数据编辑窗口中, 新建一个数据文件。按 Data→Orthogonal Design→Generate 顺序单击菜单项, 展开正交设计对话框, 见图 9-2。在 Factor Name 框中输入 a, 单击 Add 按钮, 单击 a, 单击 Define Values 按钮, 弹出 Define Values 对话框, 见图 9-3。在 Auto-Fill 下的 From 1 to 框中输入水平数 2, 单击 Fill 按钮, 在右侧的 Value 下框中自动出现 1、2, 单击 Continue 按钮, 返回正交设计对话框, 完成对 a 的定义, 依英文字母顺序逐个输入因素变量名, 单击 Define Values 按钮在相应的对话框中定义因素水平值, 直至 0 为止。

在 Data File 栏内选择 Replace Working data file, 设置将设计结果显示在工作数据文件中, 即当前的数据窗中。

选择 Reset random number seed to, 设置随机数种子, 随便填写一个正整数: 2345。随机产生正交设计方案。

单击 OK 按钮提交执行。在当前工作数据窗中生成正交表。

② 与统计书中正交表的比较

按 Data→Sort cases 顺序, 展开 Sort cases 对话框。

在左边变量名源框中, 分别选择 a、b、c、d、e、f, 将其移入到 Sort by 框中, 在 Sort order 下保持系统默认选择 Ascending (升序), 单击 OK 按钮提交执行, 则将当前数据文件中的正交表进行从小到大的有序排序, 见图 9-7, 并将它存放在 data09-02.save。

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	STATUS	CARD_
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	11
2	1	1	1	1	2	2	2	1	1	2	2	2	2	2	1	0	9
3	1	1	1	2	1	2	2	2	2	2	2	1	1	1	2	0	3
4	1	1	1	2	2	1	1	2	2	1	1	2	2	2	2	0	7
5	1	2	2	1	1	1	2	1	2	1	2	1	2	2	2	0	6
6	1	2	2	1	2	2	1	1	2	2	1	2	1	1	2	0	14
7	1	2	2	2	1	2	1	2	1	2	1	1	2	2	1	0	2
8	1	2	2	2	2	1	2	2	1	1	2	2	1	1	1	0	16
9	2	1	2	1	1	1	1	2	1	2	2	2	2	1	2	0	13
10	2	1	2	1	2	2	2	2	1	1	1	1	1	2	2	0	15
11	2	1	2	2	1	2	2	1	2	1	1	2	2	1	1	0	12
12	2	1	2	2	2	1	1	1	2	2	2	1	1	2	1	0	8
13	2	2	1	1	1	1	2	2	2	2	1	2	1	2	1	0	4
14	2	2	1	1	2	2	1	2	2	1	2	1	2	1	1	0	10
15	2	2	1	2	1	2	1	1	1	1	2	2	1	2	2	0	5
16	2	2	1	2	2	1	2	1	1	2	1	1	2	1	2	0	1

图 9-7 生成的正交设计表

将上述正交表与教科书中可查到的标准正交表进行比较后发现, 它们之间存在如下对应关系 (它不是固定不变的, 设置不同的随机数, 结果不同, 每次都需要重新核对), 见表 9-10。

表 9-10 标准正交表与本次在 SPSS16.0 中生成的正交表的对应关系

原表列号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
生成列号	a	b	c	d	h	i	o	e	l	n	m	f	j	g	k

根据表 9-10 对生成正交表的数据文件中的标签名, 用表 9-10 的对应列号进行说明, 并修改其测度水平为 Nominal (名义测度)。

在输入 1—9 时, 要在它们之前加 0, 如, 01、02、…、09。

③ 对生成的正交表进行列置换

按 Data→Sort Variables 顺序, 展开 Sort Variables 对话框, 见图 9-8。

在 Variable View Columns 下框中选择 Label, 单击 OK 按钮执行, 则在数据文件中各列按标签号的顺序得到重新排列, 此时的正交表同正交试验书中的正交表是相同的。

用表 9-9 中的表头设计内容修改数据文件中相关列的

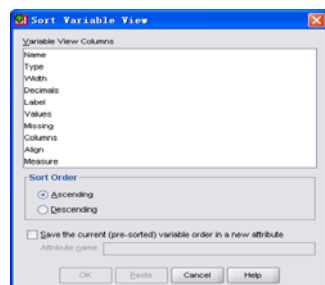


图 9-8 Sort Variables 对话框

变量名称，见图 9-9。本例已将修改完成的数据文件存放在 data09-02a.sav 中。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	STATUS_	Numeric	8	0		{0, Design}...	None	6	Right	Nominal
2	CARD_	Numeric	8	0		None	None	6	Right	Nominal
3	划臂练习	Numeric	8	0	01	{1, 1}...	None	4	Right	Nominal
4	打腿练习	Numeric	8	0	02	{1, 1}...	None	4	Right	Nominal
5	划臂练习X...	Numeric	8	0	03	{1, 1}...	None	4	Right	Nominal
6	手腿配合游	Numeric	8	0	04	{1, 1}...	None	4	Right	Nominal
7	划臂练习X...	Numeric	8	0	05	{1, 1}...	None	4	Right	Nominal
8	打腿练习X...	Numeric	8	0	06	{1, 1}...	None	3	Right	Nominal
9	o	Numeric	8	0	07	{1, 1}...	None	4	Right	Nominal
10	力量练习	Numeric	8	0	08	{1, 1}...	None	5	Right	Nominal
11	划臂练习X...	Numeric	8	0	09	{1, 1}...	None	4	Right	Nominal
12	打腿练习X...	Numeric	8	0	10	{1, 1}...	None	4	Right	Nominal
13	m	Numeric	8	0	11	{1, 1}...	None	4	Right	Nominal
14	手腿配合游...	Numeric	8	0	12	{1, 1}...	None	4	Right	Nominal
15	j	Numeric	8	0	13	{1, 1}...	None	4	Right	Nominal
16	g	Numeric	8	0	14	{1, 1}...	None	3	Right	Nominal
17	k	Numeric	8	0	15	{1, 1}...	None	3	Right	Nominal

图 9-9 正交表中各列存放的变量

在当前工作数据文件中，对系统自动赋予四个因素的值标签名“1”、“2”用具体水平对应的实际内容进行修改。如划臂练习，1=0.5KM、2=1KM。并删除 Label 中的数字。以便打印正交设计方案时能显示 4 个因素的中文变量名。

仿例 9.1 中的做法，对表 9-9 中各列的变量名及 4 个试验因素的值标签，按表 9-8 中的水平值做出定义，并将修改后的结果存放在 data09-02b.sav 中。

按 Data→Orthogonal Design→Display 顺序单击菜单项，展开 Display Design 对话框，见图 9-5。在左边变量名源框中，选择 4 个因素的变量名划臂练习、打腿练习、手臂配合练习、力量练习，将其移入到 Factors 框中，在 Format 选项中选择 Listing for experimenter，要求输出试验方案。单击 OK 按钮运行，则在输出窗口中出现所要的输出结果，见表 9-11。

表 9-11 4 因素 2 水平正交试验方案

Card List					
	Card ID	划臂练习	打腿练习	手腿配合游	力量练习
1	11	0.5KM	1KM	4KM	10组
2	9	0.5KM	1KM	4KM	20组
3	3	0.5KM	1KM	6KM	10组
4	7	0.5KM	1KM	6KM	20组
5	6	0.5KM	1.5KM	4KM	10组
6	14	0.5KM	1.5KM	4KM	20组
7	2	0.5KM	1.5KM	6KM	10组
8	16	0.5KM	1.5KM	6KM	20组
9	13	0.8KM	1KM	4KM	10组
10	15	0.8KM	1KM	4KM	20组
11	12	0.8KM	1KM	6KM	10组
12	8	0.8KM	1KM	6KM	20组
13	4	0.8KM	1.5KM	4KM	10组
14	10	0.8KM	1.5KM	4KM	20组
15	5	0.8KM	1.5KM	6KM	10组
16	1	0.8KM	1.5KM	6KM	20组

表中第一列可以理解为试验号，第 2 列为随机安排的原始条件基本一致的被试对象，

因此，第一行的含义为：第一次试验中安排被试对象 11 号，接受的训练方案为划臂练习 0.5KM、打退练习 1KM、手臂配合练习 4KM、力量练习 10 组，其余类推。

9.4 正交试验设计的极差分析

极差分析法也称直观分析法，简称 R 法，是正交试验结果分析中计算简便、形象直观、简单易懂的一种最常用的方法。

9.4.1 极差分析的基本步骤

1. 对因素所在列和交互作用所在列，计算列中各水平对应的试验结果的和 K_{jm} ，其中， j 表示第 j 列， m 表示因素或交互作用第 m 个水平。

2. 计算各因素、各交互作用的各水平的均值 $\bar{K}_{jm} = \frac{K_{jm}}{n_m}$ ，由 \bar{K}_{jm} 的大小可以判断 j 因素或 j 交互作用的优水平和各因素的优水平组合，即最优组合。

3. 计算各列的极差 $R_j = \max(K_{j1}, K_{j2}, \dots, K_{jm}) - \min(K_{j1}, K_{j2}, \dots, K_{jm})$ 。 R_j 反映了第 j 列因素或交互作用的水平变动时，试验指标的变动幅度。 R_j 越大，说明该列因素或交互作用对试验指标的影响越大，故它越重要。所以，根据 R_j 的大小可以反映出因素的主次关系。

9.4.2 极差分析法的实例分析

1. 无交互作用的情形

(1) 等水平正交试验设计单因变量的极差分析

例 9.3 以例 9.1 的正交试验的设计方案为基础，按选定的这 9 个试验进行试验，测得试验结果见表 9-12，试对其进行极差分析。

表 9-12 影响铁水温度 3 因素 3 水平的正交试验结果 (°C)

试验编号	1	2	3	4	5	6	7	8	9
铁水温度	1365	1395	1385	1390	1395	1380	1390	1390	1410

用 SPSS 进行极差分析的具体步骤如下：

① 在正交试验设计方案中录入试验结果

打开存放正交试验设计方案的 data09-01a.sav 数据文件，将 STATUS_ 变量修改为铁水温度，并在铁水温度变量中依次输入表 9-12 中的实验结果，将修改和输入试验结果的数据文件另存为 data09-01b.sav。

② 计算各因素下各水平试验结果均值的极差并选优

按 Analyze→General Linear Model→Univariate 顺序, 打开 Univariate 主对话框, 见图 8-1。

在左侧因素和协变量框中, 选择铁水温度, 并将其移入到 Dependent Variable 框中; 用同样方法分别将 A、B、D 变量移入到 Fixed Factor(s) 框中。

单击 Options 按钮, 在弹出的 Options 对话框中, 从 Factor(s) and Factor interactions 框中选择 A、B、D 变量将其移入到 Display Means for 框, 选择 Compare main effects 选项。单击 Continue 按钮, 返回 Univariate 主对话框, 单击 OK 按钮运行, 则在输出窗口中得到 11 张表, 其中 Pairwise Comparisons 的 3 张表是本例中需要的计算结果, 分别见表 9-13、表 9-14 和表 9-15。

从各表第二列均值差异列中, 可以找到均值差异均大于或等于 0 的一行, 这表示因素某个水平的均值与其他各水平的均值比较时, 该水平对应的均值为最大, 而均值差异正值中最大均值差, 正是所要求的极差。

为使我们更容易找到极差出现的行, 在表中, 我们已经用黑框作了标注。

因此, 在表 9-13 见到的因素 A 中, 极差为 15, 而这一行对应于 A 因素的第三个水平, 即 A_3 的均值最大, 所以优方案为 A_3 。

表 9-13 因素 A 各水平均值的极差

Dependent Variable: 铁水温度						
Pairwise Comparisons						
(I) A	(J) A	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference ^c	
					Lower Bound	Upper Bound
焦比1:16	焦比1:18	-6.667 ^{a,b}
焦比1:16	焦比1:14	-15.000 ^{a,b}
焦比1:18	焦比1:16	6.667 ^{a,b}
焦比1:18	焦比1:14	9.333 ^{a,b}
焦比1:14	焦比1:16	15.000 ^{a,b}
焦比1:14	焦比1:18	8.333 ^{a,b}

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

同理, 从表 9-14 中可见, 在因素 B 中, 极差为 11.67, 而这一行对应于 B 因素的第二个水平, 即 B_2 的均值最大, 所以优方案为 B_2 。

表 9-14 因素 B 各水平均值的极差

Dependent Variable: 铁水温度						
Pairwise Comparisons						
(I) B	(J) B	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference ^c	
					Lower Bound	Upper Bound
风压170	风压230	-11.667 ^{a,b}
风压170	风压200	10.000 ^{a,b}
风压230	风压170	11.667 ^{a,b}
风压230	风压200	1.667 ^{a,b}
风压200	风压170	10.000 ^{a,b}
风压200	风压230	-1.667 ^{a,b}

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

从表 9-15 中可见，在因素 C 中，极差为 20.00，而这一行对应于 C 因素的第二个水平，即 C_2 的均值最大，所以优方案为 C_2 。

表 9-15 因素 C 各水平均值的极差

Pairwise Comparisons						
Dependent Variable: 铁水温度						
		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^c	
(I) D	(J) D				Lower Bound	Upper Bound
底焦高度 1.2	底焦高度 1.5	-20.000 ^{a,b}
	底焦高度 1.3	-11.667 ^{a,b}
底焦高度 1.5	底焦高度 1.2	20.000 ^{a,b}
	底焦高度 1.3	8.333 ^{a,b}
底焦高度 1.3	底焦高度 1.2	11.667 ^{a,b}
	底焦高度 1.5	-8.333 ^{a,b}

Based on estimated marginal means
a. An estimate of the modified population marginal mean (I).
b. An estimate of the modified population marginal mean (J).
c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

以上三个优方案中，极差最大的为 C_2 等于 20.00 大于 15 和 11.67，三个因素中 C 因素的水平改变时对试验结果的影响最大。因此，因素 C 是我们需要考虑的主要因素。当它取第二水平时，对应的均值最大，故取它的第二个水平最好。极差第二大的为 A 因素，当 A 因素取第三个水平对应的均值最大，故取 A 因素的第三个水平最好。同理，在 B 因素中，应取 B 因素的二水平最好。由此，由极差分析得出的最好的方案为 $C_2A_3B_2$ 。

对照原始的试验记录，我们看到，这里给出的最好方案在已经做过试验的 9 次中没有出现，与它相近的是第 9 号试验，在第 9 号试验中只有风压 B 不是处在最好水平，且风压 B 是三个因素中影响最小的，这也说明我们找出的最好方案是符合实际的。为最终确定这个方案是不是最好的方案，可以按这个方案再做一次试验，同第 9 号试验的结果铁水温度 1410℃ 进行比较，以便确定是 $C_2A_3B_2$ 为最好方案（若铁水温度高于 1410℃），还是 $C_2A_3B_3$ 为最好方案（若铁水温度低于 1410℃）。

(2) 混合水平正交试验设计单因变量的极差分析

例 9.4 某农科站对晚稻品种和栽培措施进行试验。试验指标为晚稻的产量，评价标准为产量越高越好，根据以往的经验，选取影响产量的主要因素 5 个，各因素名称及其水平见表 9-16。在不考虑交互作用的前提下，设计混合水平正交试验方案，按正交试验方案进行实施，对测得的试验结果进行分析，选出最好的生产方案。

表 9-16 影响晚稻产量 5 个因素的各水平值

因素 水平	品种 A	栽种规格 B	每穴株树 C	追肥量 D (斤/亩)	穗肥量 E (斤/亩)
1	甲	4×3	7~8	15	3
2	乙	4×4	4~5	20	0
3	丙				
4	丁				

①设计 1 因素 4 水平和 4 因素 2 水平的混合水平的正交试验设计方案

由于本例不需要考虑因素间的交互作用，因此，加上一个误差的空白列，共需要 6 列的正交表，把表 9-16 中的 5 个因素依次放在前 5 列(A—E)，把误差项放在第 6 列(F)，则利用例 9.1 中介绍的方法，可在 SPSS 中生成 5 因素混合水平的正交表，见 data09-03.sav，然后定义各因素的水平值和标签名后，在输出窗口中生成表 9-17 所示的 5 因素混合水平的正交试验设计方案，并将其另存为 data09-03a.sav。表 9-17 表明，共需要做 16 次试验，每行中各因素水平的组合即是每次试验的方案。

表 9-17 影响晚稻产量的 5 因素混合水平的正交试验设计方案

Card List						
	Card ID	品种	栽种规格	每穴株数	追肥量	穗肥量
1	14	甲	4×3	7~8	15	3
2	11	甲	4×3	7~8	15	0
3	6	甲	4×4	4~5	20	3
4	8	甲	4×4	4~5	20	0
5	16	乙	4×3	4~5	20	3
6	3	乙	4×3	4~5	20	0
7	10	乙	4×4	7~8	15	3
8	7	乙	4×4	7~8	15	0
9	13	丙	4×3	4~5	15	3
10	4	丙	4×3	4~5	15	0
11	9	丙	4×4	7~8	20	3
12	15	丙	4×4	7~8	20	0
13	12	丁	4×3	7~8	20	3
14	1	丁	4×3	7~8	20	0
15	5	丁	4×4	4~5	15	3
16	2	丁	4×4	4~5	15	0

②试验结果

通过严格按以上方案组织实施试验，测得 16 次试验所得的产量（斤）依次为 694，664，714，650，650，646，670，652，646，600，630，670，660，670，670，650。

在 SPSS 中，打开 data09-03a.sav，删除 STATUS_变量，新增数字型产量变量，将上述 16 次试验结果依次录入到 data09-03a.sav 的产量变量中，并将修改后的数据文件另存为 data09-03b.sav 中。

③计算各因素下各水平试验结果均值的极差并选优

按 Analyze→General Linear Model→Univariate 顺序，打开 Univariate 主对话框，见图 8-1。

在左侧因素和协变量框中，选择产量变量，单击向右箭头，将其送入 Dependent Variable 框中；用同样方法分别将 A、B、C、D、E 变量送入 Fixed Factor(s) 框中。

单击 Options 按钮，在弹出的 Options 对话框中，从 Factor(s) and Factor interactions 框中选择 A、B、C、D、E 变量将其移入到 Display Means for 框中，选择 Compare main effects 选项。单击 Continue 按钮，返回 Univariate 主对话框，单击 OK 按钮运行，则在输出窗口中得到 17 张表，其中 Pairwise Comparisons 的 5 张表是本例中需要的计算结果，分别见表 9-18、表 9-19、表 9-20、表 9-21 和表 9-22。

极差寻找和因素最好水平的确定方法同上例。

表 9-18 因素 A 各水平平均值的极差

Pairwise Comparisons						
Dependent Variable: 产量						
(I) 品种	(J) 品种	Mean Difference (I-J)	Std. Error	Sig. ^c	95% Confidence Interval for Difference ^c	
					Lower Bound	Upper Bound
甲	乙	26.000 ^{a,b}
	丙	44.000 ^{a,b}
	丁	18.000 ^{a,b}
乙	甲	-26.000 ^{a,b}
	丙	18.000 ^{a,b}
	丁	-8.000 ^{a,b}
丙	甲	-44.000 ^{a,b}
	乙	-18.000 ^{a,b}
	丁	-26.000 ^{a,b}
丁	甲	18.000 ^{a,b}
	乙	8.000 ^{a,b}
	丙	26.000 ^{a,b}

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-19 因素 B 各水平平均值的极差

Pairwise Comparisons						
Dependent Variable: 产量						
(I) 栽种规格	(J) 栽种规格	Mean Difference (I-J)	Std. Error	Sig. ^c	95% Confidence Interval for Difference ^c	
					Lower Bound	Upper Bound
4X3	4X4	-9.500 ^{a,b}
4X4	4X3	9.500 ^{a,b}

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-20 因素 C 各水平平均值的极差

Pairwise Comparisons						
Dependent Variable: 产量						
(I) 每穴株数	(J) 每穴株数	Mean Difference (I-J)	Std. Error	Sig. ^c	95% Confidence Interval for Difference ^c	
					Lower Bound	Upper Bound
7~8	4~5	10.500 ^{a,b}
4~5	7~8	-10.500 ^{a,b}

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-21 因素 D 各水平平均值的极差

Pairwise Comparisons						
Dependent Variable: 产量						
(I) 追肥量	(J) 追肥量	Mean Difference (I-J)	Std. Error	Sig. ^c	95% Confidence Interval for Difference ^c	
					Lower Bound	Upper Bound
15	20	-5.500 ^{a,b}
20	15	5.500 ^{a,b}

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-22 因素 E 各水平均值的极差

Dependent Variable: 产量						
Pairwise Comparisons		Mean Difference (I- J)	Std. Error	Sig. ^c	95% Confidence Interval for Difference ^c	
(I) 穗 肥量	(J) 穗 肥量				Lower Bound	Upper Bound
3	0	16.500 ^{a,b}
0	3	-16.500 ^{a,b}

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

从表 9-18 可见, 因素 A 中品种甲 (第一个水平) 均值最大, 极差为 44.00, 所以优方案为 A_1 。从表 9-19 可见, 因素 B 中栽种规格 4×4 (第二个水平) 均值最大, 极差为 9.500, 所以优方案为 B_2 。从表 9-20 可见, 因素 C 中每穴株数 7~8 (第一个水平) 均值最大, 极差为 10.50, 所以优方案为 C_1 。从表 9-21 可见, 因素 D 中追肥量 20 (第二个水平) 均值最大, 极差为 5.50, 所以优方案为 D_2 。从表 9-22 可见, 因素 E 中穗肥量 3 (第一个水平) 均值最大, 极差为 16.500, 所以优方案为 E_1 。

以上五个优方案中, 极差最大的为 A_1 等于 44.00, 因此, 五个因素中 A 因素的水平改变时对试验结果的影响最大。因此, 因素 A 是我们需要考虑的主要因素。当它取第一水平时, 对应的均值最大, 故取它的第一个水平最好。极差第二大的为 E 因素, 当 E 因素取第一个水平对应的均值最大, 故取 E 因素的第一个水平最好。极差第三大的为 C 因素, 当 C 因素取第一个水平对应的均值最大, 故取 C 因素的第一个水平最好。同理, 在 B 因素中, 应取 B 因素的第二水平最好; 在 D 因素中, 应取 D 因素的第二水平最好。由此, 由极差分析得出的最好的方案为 $A_1E_1C_1B_2D_2$ 。

对照原始的试验记录, 我们看到, 这里给出的最好方案在已经做过试验的 16 次中没有出现, 与它相近的是第 3 号试验, 在第 3 号试验中的方案为 $A_1E_1C_2B_2D_2$, 两者相比, 在第三重要因素上有区别, 为最终确定这个方案是不是最好的方案, 可以按这个方案再做一次试验, 然后同第 3 号试验的结果产量 714 斤进行比较, 如果试验结果产量大于 714 斤, 说明 $A_1E_1C_1B_2D_2$ 为最好方案, 否则, 认为 $A_1C_2D_2B_2E_1$ 是最好方案。

(3) 等水平正交试验设计多因变量的极差分析

① 综合平衡法

在因变量为多个时, 一般先分别考察每个因素对各因变量的影响, 再进行综合分析比较, 确定出最好的水平, 最后得出最好试验方案的方法称为综合评定法。

例 9.5 为提高某产品的质量, 对生产该产品的原料进行配方试验。用检验抗压强度 (公斤/个)、落下强度 (0.5 米/次) 和裂纹度三项指标来评价该产品的质量, 前两个指标值越大越好, 最后一个指标值越小越好。由以往的经验知道, 配方中有三个重要因素: 水分 (%)、粒度 (%) 和碱度。它们各有三个水平, 见表 9-23。试验安排和试验结果见

表 9-24。试用极差分析中综合平衡法进行实验分析，找出最好的配方方案。（数据来源：吉林教育出版社，1987 年 9 月第一版，清华大学数学系概率统计教研组编《概率论与数量统计》，P347）

表 9-23 因素及其水平

因素 水平	A 水分	B 粒度	C 碱度
1	8	4	1.1
2	9	6	1.3
3	7	8	1.5

表 9-24 试验方案及试验结果

列号 试验号	1 A	2 B	3 C	各指标的试验结果		
				抗压强度	落下强度	裂纹度
1	1	1	1	11.5	1.1	3
2	1	2	2	4.5	3.6	4
3	1	3	3	11.0	4.6	4
4	2	1	2	7.0	1.1	3
5	2	2	3	8.0	1.6	2
6	2	3	1	18.5	15.1	0
7	3	1	3	9.0	1.1	3
8	3	2	1	8.0	4.6	2
9	3	3	2	13.4	20.2	1

1) 在 SPSS 中建立数据文件，见 data09-04.sav。

2) 计算各因素下各水平试验结果均值的极差并选优

按 Analyze→General Linear Model→Multivariate 顺序，打开 Multivariate 主对话框。

在左侧因素和协变量框中，选择抗压强度、落下强度、裂纹度变量，单击向右箭头，将其送入 Dependent Variable 框；用同样方法分别将 A、B、C 变量送入 Fixed Factor (s) 框中。

单击 Options 按钮，在弹出的 Options 对话框中，从 Factor(s) and Factor interactions 框中选择 A、B、C 变量将其移入到 Display Means for 框中，选择 Compare main effects 选项。单击 Continue 按钮返回 Multivariate 主对话框，单击 OK 按钮运行，则在输出窗口中得到 15 张表，其中 Pairwise Comparisons 的 3 张表是本例中需要的计算结果，分别见表 9-25、表 9-26 和表 9-27。

这些表是单因变量的复合表，极差寻找和因素最好水平的确定方法同上例。

表 9-25 因素 A 在三个因变量上各水平平均值的极差

Pairwise Comparisons						
Dependent Variable	(I) A	(J) A	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference ^a
						Lower Bound
抗压强度	1	2	-2.167 ^{**}	.	.	.
		3	-1.133 ^{**}	.	.	.
	2	1	2.167 ^{**}	.	.	.
		3	1.033 ^{**}	.	.	.
3	1	1.133 ^{**}	.	.	.	
	2	-1.033 ^{**}	.	.	.	
落下强度	1	2	-2.833 ^{**}	.	.	.
		3	-5.533 ^{**}	.	.	.
	2	1	2.833 ^{**}	.	.	.
		3	-2.700 ^{**}	.	.	.
3	1	5.533 ^{**}	.	.	.	
	2	2.700 ^{**}	.	.	.	
裂紋度	1	2	2.000 ^{**}	.	.	.
		3	1.667 ^{**}	.	.	.
	2	1	-2.000 ^{**}	.	.	.
		3	-.333 ^{**}	.	.	.
3	1	-1.667 ^{**}	.	.	.	
	2	.333 ^{**}	.	.	.	

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-26 因素 B 在三个因变量上各水平平均值的极差

Pairwise Comparisons							
Dependent Variable	(I) B	(J) B	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^c	
						Lower Bound	Upper Bound
抗压强度	1	2	2.333 ^{**}
		3	-5.133 ^{**}
	2	1	-2.333 ^{**}
		3	-7.467 ^{**}
落下强度	1	2	5.133 ^{**}
		3	7.467 ^{**}
	2	1	-5.133 ^{**}
		3	-12.200 ^{**}
裂纹度	1	2	-2.167 ^{**}
		3	-10.033 ^{**}
	2	1	12.200 ^{**}
		3	10.033 ^{**}

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-27 因素 C 在三个因变量上各水平平均值的极差

Pairwise Comparisons							
Dependent Variable	(I) C	(J) C	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^c	
						Lower Bound	Upper Bound
抗压强度	1	2	4.367**
		3	3.333**
	2	1	-4.367**
		3	-1.033**
落下强度	1	2	-3.333**
		3	1.033**
	2	1	3.333**
		3	5.867**
裂纹度	1	2	-4.500**
		3	-5.867**
	2	1	4.500**
		3	1.000**
3	1	1.333**	
	2	1.333**	

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

从表 9-25 可见, A 因素在三个因变量上, 最优方案分别为 A_2 、极差为 2.167, A_3 、极差为 5.533, A_2 、极差为 -2.000 (越小越好)。

从表 9-26 可见, B 因素在三个因变量上, 最优方案分别为 B_3 、极差为 7.467, B_3 、极差为 12.200, B_3 、极差为 -1.333 (越小越好)。

从表 9-27 可见, C 因素在三个因变量上, 最优方案分别为 C_1 、极差为 4.367, C_2 、极差为 5.867, C_1 、极差为 -1.333 (越小越好)。

综上所述, 根据单因变量的评价标准可知, 对于抗压强度指标最好的方案为: $B_3C_1A_2$; 对落下强度指标最好的方案是: $B_3C_2A_3$; 对于裂纹度指标最好的方案是: $A_2B_3C_1$ 。很明显, 这三个方案不完全相同。

对于粒度 B 而言, 在对抗强度和落下强度两个指标上它是影响最大的因素。在三个指标上都是取第三个水平时, 得到最大极差。因而, 粒度取第三个水平 8 为最好。

对于碱度 C 而言, 在三个指标上, 碱度 C 都不是最主要的, 它是次要因素, 对抗压强度和裂纹度指标, 碱度 C 取一水平碱度为 1.1 时最好, 对落下强度而言, 碱度取 1.3 为最好, 但取一水平碱度为 1.1 时, 其均值和 1.3 时的均值相差不是很大, 故综合考虑碱度取一水平为好。

对于水分 A 而言, 在裂纹度指标上水分的极差最大, 即水分是影响最大的因素, 而在抗压强度和落下强度指标上, 它是影响最小的指标, 故应主要考虑它对裂纹度的影响, 对裂纹度, 水分取第二个水平时能使裂纹度的均值达到最小, 故取第二个水平 9 为最好。

因此, 较好的试验方案应为: $B_3C_1A_2$, 即粒度 8、碱度 1.1 和水分 9。

② 综合评分法

当我们能对因变量的结果转换成分值, 并能给出因变量之间重要性的数量关系时, 可以采用综合评分法来进行分析, 即相当于又回到了单因变量的情形, 这可以使得分析工作变得相对简单些。

综合评分法是将多个因变量的试验结果先转换成分值, 然后根据各因变量重要性的大小, 来确定每个因变量的权重, 再进行加权处理计算总分, 这种用分数将多个因变量变成一个因变量再进行分析的方法, 称为综合评分法。

例 9.6 某厂生产一种化工产品, 需要检验两个指标: 核酸纯度和回收率, 这两个指标都是越高越好, 根据经验知道, 这两个指标的重要性是不相同的, 纯度的重要性比回收率的重要性大, 相当于 4 倍的关系。影响这两个指标的因素主要有三个, 现各取三个水平, 见表 9-28, 采用 $L_9(3^4)$ 正交表安排试验, A、B 安排在前两列, C 安排在第 4 列, 试验方案和结果见表 9-29。试用极差分析中的综合评分法找出相对较好的方案, 使这两个指标值都有提高。

表 9-28 三个因素的水平值

因素 水平	A 时间 (小时)	B 加料中核酸含量	C 加水量
1	25	7.5	1:6
2	5	9.0	1:4
3	1	6.0	1:2

表 9-29 试验方案与试验结果

列号 试验号	A	B	C	各指标试验结果	
				纯度	回收率
1	1	1	1	17.5	30.0
2	1	2	2	12.0	41.2
3	1	3	3	6.0	60.0
4	2	1	3	8.0	24.2
5	2	2	1	4.5	51.0
6	2	3	2	4.0	58.4
7	3	1	2	8.5	31.0
8	3	2	3	7.0	20.5
9	3	3	1	4.5	73.5

1) 在 SPSS 中建立数据文件, 见 data09-05.sav。

2) 计算综合评分

由于纯度指标和回收率指标都是越高越好, 因此, 可以直接用两个指标的测试值看成是其分值, 又因为纯度的重要性相当于回收率的 4 倍, 故可以构造综合评分的计算公式为

$$\text{综合评分} = 4 \times \text{纯度得分} + \text{回收率得分}$$

按 Transform→Compute 顺序打开 Compute 对话框, 见图 2-53。在 Target Variable 的下框中输入目标变量名综合评分。在 Numeric Expression 的下框中输入 4*纯度+回收率。单击 OK 按钮运行, 则在当前工作的数据文件中, 出现变量名为综合评分的新变量及其计算值, 将存有综合评分新变量及其计算值的数据文件另存为 data09-05a.sav 中。

现在, 用综合得分作为因变量, 则本例就回到了前面做过的单因变量解题模式。

3) 计算各因素下各水平试验结果均值的极差并选优

仿例 9.3 中的做法, 可得表 9-30、表 9-31 和表 9-32。

从表 9-30 可见, A_1 为优, 此时, 极差为 24.522。

从表 9-31 可见, B_3 为优, 此时, 极差为 14.400。

从表 9-32 可见, C_1 为优, 此时, 极差为 23.933。

综上所述, 极差最大的为 A 因素, 它是影响最大的因素, 其次为 C 因素, 极差最小的为 B 因素, 所以, 综合考虑最好的试验方案为 $A_1C_1B_3$, 即

表 9-30 因素 A 各水平平均值的极差

Pairwise Comparisons

Dependent Variable: 综合评分

(I) 时间	(J) 时间	Mean Difference (I-J)	Std. Error	Sig. ^c	95% Confidence Interval for Difference ^c	
					Lower Bound	Upper Bound
25	5	24.533 ^{a,b}
	1	22.733 ^{a,b}
5	25	-24.533 ^{a,b}
	1	-1.800 ^{a,b}
1	25	-22.733 ^{a,b}
	5	1.800 ^{a,b}

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-31 因素 B 各水平平均值的极差

Pairwise Comparisons

Dependent Variable: 综合评分

(I) 核酸含量	(J) 核酸含量	Mean Difference (I-J)	Std. Error	Sig. ^c	95% Confidence Interval for Difference ^c	
					Lower Bound	Upper Bound
7.5	9.0	4.833 ^{a,b}
	6.0	-9.567 ^{a,b}
9.0	7.5	-4.833 ^{a,b}
	6.0	-14.400 ^{a,b}
6.0	7.5	9.567 ^{a,b}
	9.0	14.400 ^{a,b}

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-32 因素 C 各水平平均值的极差

Pairwise Comparisons

Dependent Variable: 综合评分

(I) 加水量	(J) 加水量	Mean Difference (I-J)	Std. Error	Sig. ^c	95% Confidence Interval for Difference ^c	
					Lower Bound	Upper Bound
1:6	1:4	10.633 ^{a,b}
	1:2	23.933 ^{a,b}
1:4	1:6	-10.633 ^{a,b}
	1:2	13.300 ^{a,b}
1:2	1:6	-23.933 ^{a,b}
	1:4	-13.300 ^{a,b}

Based on estimated marginal means

a. An estimate of the modified population marginal mean (I).

b. An estimate of the modified population marginal mean (J).

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

A₁ 时间 第 1 个水平 25 小时

C₁ 加水量 第 1 个水平 1：6

B₃ 料中核酸含量 第 3 个水平 6.0

由于这个方案在试验中没有做过，所以，可以按此方案再做一次试验，以此用实践来检验所得到的方案是否是真正最好的方案。

2. 有交互作用的情形

例 9.7 以例 9.2 中的正交试验设计方案为基础, 经半年训练后, 测得各试验方案下的 800 米自由泳成绩提高值分别为, 4.40, 4.50, 9.10, 8.30, 7.80, 3.50, 9.60, 8.70, 5.90, 6.70, 4.40, 4.60, 9.40, 8.90, 7.80, 8.80, 这些试验结果已经存在在数据文件 data09-02c.sav 的试验结果变量中, 试对其进行极差分析, 找出最好方案。

【题析】 在有交互作用的情形下, 只要把交互作用当成一个新的因素看待即可, 则对其进行极差分析的方法就同以上的例中没有多大区别了。

在 SPSS 中的解题步骤如下:

计算各因素及交互作用下各水平试验结果均值的极差并选优

同例 9.3 中的做法相似, 可得表 9-33、表 9-34、表 9-35、表 9-36、表 9-37、表 9-38、表 9-39、表 9-40、表 9-41、表 9-42。

由于各因素都是二水平, 因此, 两个水平之间的差值就是所要求的极差。

表 9-33 划臂练习因素各水平均值的极差

Pairwise Comparisons

Dependent Variable: 试验结果

(I) 划臂练习	(J) 划臂练习	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
0.5KM	0.8KM	-.075	.487	.884	-1.327	1.177
0.8KM	0.5KM	.075	.487	.884	-1.177	1.327

Based on estimated marginal means

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-34 打腿练习因素各水平均值的极差

Pairwise Comparisons

Dependent Variable: 试验结果

(I) 打腿练习	(J) 打腿练习	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
1KM	1.5KM	-2.075 [*]	.487	.008	-3.327	-.823
1.5KM	1KM	2.075 [*]	.487	.008	.823	3.327

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-35 划臂与打腿练习交互作用各水平均值的极差

Pairwise Comparisons

Dependent Variable: 试验结果

(I) 划臂练习×打腿练习	(J) 划臂练习×打腿练习	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
1	2	1.250	.487	.050	-.002	2.502
2	1	-1.250	.487	.050	-2.502	.002

Based on estimated marginal means

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-36 手腿配合游因素各水平平均值的极差

Pairwise Comparisons

Dependent Variable: 试验结果

(I) 手腿配合游	(J) 手腿配合游	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
4KM	6KM	-1.275 [*]	.487	.047	-2.527	-.023
6KM	4KM	1.275 [*]	.487	.047	.023	2.527

Based on estimated marginal means

^a. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-37 划臂练习与手腿配合游交互作用各水平平均值的极差

Pairwise Comparisons

Dependent Variable: 试验结果

(I) 划臂练习X手腿配合游	(J) 划臂练习X手腿配合游	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
1	2	-2.600 [*]	.487	.003	-3.852	-1.348
2	1	2.600 [*]	.487	.003	1.348	3.852

Based on estimated marginal means

^a. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-38 打腿练习与手腿配合游交互作用各水平平均值的极差

Pairwise Comparisons

Dependent Variable: 试验结果

(I) 打腿练习X手腿配合游	(J) 打腿练习X手腿配合游	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
1	2	.050	.487	.922	-1.202	1.302
2	1	-.050	.487	.922	-1.302	1.202

Based on estimated marginal means

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-39 力量练习因素各水平平均值的极差

Pairwise Comparisons

Dependent Variable: 试验结果

(I) 力量练习	(J) 力量练习	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
10组	20组	.550	.487	.310	-.702	1.802
20组	10组	-.550	.487	.310	-1.802	.702

Based on estimated marginal means

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-40 划臂练习与力量练习交互作用各水平平均值的极差

Pairwise Comparisons

Dependent Variable: 试验结果						
(I) 划臂练习×力量练习	(J) 划臂练习×力量练习	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
1	2	.925	.487	.116	-.327	2.177
2	1	-.925	.487	.116	-2.177	.327

Based on estimated marginal means

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-41 打腿练习与力量练习因素各水平平均值的极差

Pairwise Comparisons

Dependent Variable: 试验结果						
(I) 打腿练习×力量练习	(J) 打腿练习×力量练习	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
1	2	-.625	.487	.256	-1.877	.627
2	1	.625	.487	.256	-.627	1.877

Based on estimated marginal means

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-42 手臂配合游与力量练习交互作用各水平平均值的极差

Pairwise Comparisons

Dependent Variable: 试验结果						
(I) 手臂配合游×力量练习	(J) 手臂配合游×力量练习	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
1	2	.425	.487	.423	-.827	1.677
2	1	-.425	.487	.423	-1.677	.827

Based on estimated marginal means

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-43 极差分析表

因素	极差	优方案
划臂练习	0.75	2 水平
打腿练习	2.075	2 水平
划臂练习×打腿练习	1.250	1 水平
手腿配合游	1.225	2 水平
划臂练习×手腿配合游	2.600	2 水平
打腿练习×手腿配合游	0.050	1 水平
力量练习	0.550	1 水平
划臂练习×力量练习	0.925	1 水平
打腿练习×力量练习	0.625	2 水平
手腿配合游×力量练习	0.425	1 水平

为更加清楚地看清上述表中的内容，我们将它整理成表 9-42 所示的极差分析表。

从表 9-43 中的极差列可见，各行的极差是不同的，这说明各因素的水平改变时对试验指标的影响是不同的。极差越大，说明这个因素的水平改变时对试验指标的影响越大。极差最大的那个因素的水平改变时对试验指标的影响最大，那个因素就是我们要考虑的主要因素。

从表 9-42 中的极差大小可以看出，影响最大的是划臂练习×手腿配合游，以 2 水平为好，其次是打腿练习，以 2 水平为好，第

三是划臂练习×打腿练习，以1水平为好，第四为手腿配合游，以2水平为好，其他的极差很小，对试验影响很小，可以忽略不计。综合分析考虑，最好的方案应当是打腿练习2手腿配合游2和划臂练习2。由于现在安排的力量练习对试验影响很小，可以不考虑。

故最佳搭配方案为：打腿练习1.5千米+手腿配合游6千米+划臂练习0.8千米。

9.5 正交试验设计的方差分析

极差分析虽然简单易行，但它不能区分数据的波动到底是由试验条件的改变引起的，还是由试验误差引起的，因此，它不能提供试验的精度，以及不能给出各因素对试验结果的重要程度的精确的数量估计。对试验结果进行方差分析可以弥补极差分析的不足。

9.5.1 正交试验设计方差分析的基本原理

1. 偏差平方和的分解

在正交试验中，总的偏差平方和与总的自由度为

$$SS_T = \sum_{i=1}^n (x_i - \bar{x})^2, \quad f_T = n - 1$$

各列的偏差平方和与自由度为

$$SS_j = r \sum_{i=1}^m (\bar{K}_{ij} - \bar{x})^2 \quad (j=1, 2, \dots, k), \quad f_j = m - 1$$

误差平方和与自由度为

$$SS_e = \sum_{k_{\text{空}}} SS_j, \quad f_e = \sum_{k_{\text{空}}} f_j$$

可以证明

$$\begin{aligned} SS_T &= \sum_{j=1}^k SS_j = \sum_{k_{\text{因}}} SS_j + \sum_{k_{\text{交}}} SS_j + \sum_{k_{\text{空}}} SS_j \\ f_T &= \sum_{j=1}^k f_j = \sum_{k_{\text{因}}} f_j + \sum_{k_{\text{交}}} f_j + \sum_{k_{\text{空}}} f_j \end{aligned}$$

式中， $k_{\text{因}}$ ， $k_{\text{交}}$ ， $k_{\text{空}}$ 分别为试验因素、试验考虑的交互作用和正交表中的空列的列数。

2. 显著性检验

首先计算均方差：

$$S_j^2 = \frac{SS_j}{f_j}, \quad S_{\text{因}}^2 = \frac{SS_{\text{因}}}{f_{\text{因}}}, \quad S_{\text{交}}^2 = \frac{SS_{\text{交}}}{f_{\text{交}}}, \quad S_e^2 = \frac{SS_e}{f_e}$$

所要检验的原假设 H_0 ：某因素或某交互作用不显著。

在原假设成立时，可以证明，统计量

$$F = \frac{S_{\text{因}}^2 (\text{或 } S_{\text{交}}^2)}{S_e^2} \sim F(f_{\text{因}} (\text{或 } f_{\text{交}}), f_e)。$$

所以，根据 F 分布，可以检验原假设成立的概率。

9.5.2 正交试验设计方差分析实例

1. 不考虑交互作用的情形

试验设计方案可直接用 SPSS16.0 提供的正交设计程序产生，获得试验结果后，可以直接用 SPSS 中对一般多因素进行方差分析的方法来进行正交试验设计的方差分析。

例 9.8 试对存放在 data09-01b.sav 中的例 9.3 中取得的正交试验结果进行方差分析。其操作步骤如下：

(1) 在 SPSS 数据编辑窗口中，打开数据文件 data09-01b.sav。

(2) 定义模型

按 Analyze→General Linear Model→Univariate 顺序，展开 Univariate 主对话框，见图 8-1。在左侧变量名源框中，选中所有因素的变量名 A、B、D，将其移入到 Fixed Factor(s) 下框中，同样选中铁水温度，将其移入到 Dependent Variable 下框中。

单击 Model 按钮，展开 Model 对话框，见图 8-2。在 Specify Model 选项中选择 Custom，在 Factors & Covariates 下框中同时选中 A、B、D，在 Build Term(s) 的 Type 的下拉式选项

中，选择 Main effects，单击右移箭头按钮，将上述三个因素移入到 Model 下框中。单击 Continue 按钮，返回 Univariate 主对话框。

(3) 对方差分析的结果进行分析

单击 OK 按钮，在输出窗口中出现方差分析结果，见表 9-44。

从表 9-44 可见，在 A 因素、B 因素、C 因素（在 D 列）对铁水温度没有影响的原假设下，出现目前统计量的值或者

更极端值的双侧检验概率为均小于 0.05，故均拒绝原假设，而认为 A 因素（焦比）、B 因素（风压）、C 因素（底焦高度）对铁水温度的提高有显著性意义。

根据均方差的大小可知，因素作用的主次顺序为 CAB。

2. 考虑交互作用的情形

正交设计方案是建立在统计书中提供的正交表结合其对应的交互作用表，在正确做好表头设计后得到。一般是不能用统计书中的交互作用表结合 SPSS 正交设计程序产生的

表 9-44 方差分析表

Tests of Between-Subjects Effects

Dependent Variable: 铁水温度					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1183.333 ^a	6	197.222	71.000	.014
Intercept	1.736E7	1	1.736E7	6.250E6	.000
A	338.889	2	169.444	61.000	.016
B	238.889	2	119.444	43.000	.023
D	605.556	2	302.778	109.000	.009
Error	5.556	2	2.778		
Total	1.736E7	9			
Corrected Total	1188.889	8			

a. R Squared = .995 (Adjusted R Squared = .981)

正交表（不做任何初等转换）来安排试验的。

例 9.9 试对存放在 data09-02c.sav 中的例 9.7 中的正交试验结果进行方差分析。

【题析】 例 9.7 中的正交试验结果是建立在例 9.2 中的正交试验设计方案基础上取得的，其中用到的正交表虽然表面上也是由 SPSS 正交设计程序产生的，但它是在对照统计书中的标准正交表，经过初等转换后形成与书中正交表完全一样的正交表。所以，与其说它是由 SPSS 正交设计程序产生的正交表，倒不如说是直接用统计书中的正交表来安排试验方案的。

对它进行方差分析时的操作步骤如下：

(1) 在 SPSS 数据编辑窗口中，打开数据文件 data09-02c.sav。

(2) 定义模型

按 Analyze→General Linear Model→Univariate 顺序，展开 Univariate 对话框，见图 8-1。在左侧变量名源框中，选中所有因素的变量名（本例中为划臂练习、打腿练习、手腿配合游、力量练习），将其移入到 Fixed Factor(s)下框中，同样选中试验结果，将其移入到 Dependent Variable 下框中。单击 Model 按钮，展开 Model 对话框。在 Specify Model 选项中选择 Custom，在 Factors & Covariates 下框中同时选中划臂练习、打腿练习、手腿配合游、力量练习，在 Build Term(s)的 Type 的下拉式选项中，选择 Main effects，单击右移箭头按钮，将上述 4 个因素移入到 Model 下框中，再在 Factors & Covariates 下框中同时选中划臂练习、打腿练习、手腿配合游、力量练习，在 Build Term(s)的 Type 的下拉式选项中，选择 All 2-Way，单击右移箭头按钮，将四个因素两两间的 6 个交互左右移入到 Model 下框中。单击 Continue 按钮，返回 Univariate 对话框。

(3) 对方差分析的结果进行分析

按 OK 按钮，在输出窗口中出现方差分析结果，见表 9-45。

如果在(2)模型定义中，在 Build Term(s)的 Type 的下拉式选项中，选择 Main effects，将 4 个因素和 6 个交互作用项全部选中，并移入到 Model 下框中，返回对话框后，按 OK 运行后，则在输出窗口中得到如表 9-46 所示的方差分析的计算结果。

对这两个表对比后发现，这两张表是一致的。但如果使用交互作用表安排表头设计后直接用到由 SPSS 正交设计程序产生的正交表中，则使用上述两种定义模型的方式会得到不同的计算结果。这说明书中提供的交互作用表和正交表是配套的，不能混用。

另外，还要注意，本例各因素的水平都是 2 水平的，因此，在本例中，两种做法的结果是一致的，但需要指出的是，在多因素 3 水平或 3 水平以上时，用第二种做法计算的两个因素的交互作用是分散在多个交互作用的列变量上的，还需要合并，比较麻烦，而用第一种做法，计算到的是两个变量的各水平合并的交互作用。

从表 9-45 或表 9-46 可以知道：手腿配合游 ($P=0.047$)、打腿练习 ($P=0.008$)、划臂练习与手腿配合游的交互作用 ($P=0.003$)、划臂练习与打腿练习的交互作用 ($P=0.050$)

对 800 米成绩的提高的影响有显著性意义。其他因素和交互作用对 800 米成绩的提高的影响没有显著性意义。

表 9-45 方差分析表

Tests of Between-Subjects Effects					
Dependent Variable: 试验结果					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	63.965 ^a	10	6.397	6.740	.024
Intercept	789.610	1	789.610	832.044	.000
划臂练习	.022	1	.022	.024	.884
打腿练习	17.222	1	17.222	18.148	.008
手腿配合游	6.503	1	6.503	6.852	.047
力量练习	1.210	1	1.210	1.275	.310
划臂练习 * 力量练习	3.423	1	3.423	3.606	.116
划臂练习 * 手腿配合游	27.040	1	27.040	28.493	.003
划臂练习 * 打腿练习	6.250	1	6.250	6.586	.050
手腿配合游 * 力量练习	.723	1	.723	.761	.423
打腿练习 * 力量练习	1.562	1	1.562	1.646	.256
打腿练习 * 手腿配合游	.010	1	.010	.011	.922
Error	4.745	5	.949		
Total	858.320	16			
Corrected Total	68.710	15			

a. R Squared = .931 (Adjusted R Squared = .793)

表 9-46 方差分析表

Tests of Between-Subjects Effects					
Dependent Variable: 试验结果					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	63.965 ^a	10	6.397	6.740	.024
Intercept	789.610	1	789.610	832.044	.000
划臂练习	.022	1	.022	.024	.884
打腿练习	17.222	1	17.222	18.148	.008
手腿配合游	6.503	1	6.503	6.852	.047
力量练习	1.210	1	1.210	1.275	.310
划臂练习X打腿练习	6.250	1	6.250	6.586	.050
划臂练习X手腿配合游	27.040	1	27.040	28.493	.003
打腿练习X手腿配合游	.010	1	.010	.011	.922
划臂练习X力量练习	3.423	1	3.423	3.606	.116
打腿练习X力量练习	1.562	1	1.562	1.646	.256
手腿配合游X力量练习	.723	1	.723	.761	.423
Error	4.745	5	.949		
Total	858.320	16			
Corrected Total	68.710	15			

a. R Squared = .931 (Adjusted R Squared = .793)

第 10 章 相关与回归分析

在科学研究中,研究者经常会遇到一些同处于一个统一体中的变量。这些变量是相互联系、相互制约的。当这些变量间存在明确的因果关系时,只要知道哪个是因及因所处的状态,就能明确知道结果所处的水平。这类变量的关系是确定性的,用数学概念来讲,也就是,这些变量间存在函数关系,此时可以用一个变量的值去精确地推测另一个变量的值。例如,匀速直线运动中,距离和运动持续时间及速度的关系是确定性的,因此距离可用函数 $S = V \times T$ 精确地去描述。但在更多的实际问题中,由于许多不可控因素的存在,使得变量间的关系变得非常复杂,虽然它们之间也存在某种数量上的关系,但相互间确定性的关系,却无法轻易地用函数来表示它们之间的关系。例如,在 2008 年北京奥运会上,我们看到百米成绩好的运动员,在 200 米比赛中也有不俗的表现;又如高考模拟考试中,成绩好的学生在正式高考中一般也会有不错的高考成绩等。在对身高的研究中,人们还会发现父代身高高的,子代一般也有较高的身高,出现子代向父代的“回归”。因此,在对有一定数量关系,但不完全是因果关系的多变量之间关系的研究中,要用到本章所涉及的内容。

本章主要介绍对每个研究对象同时观测两个或多个指标的成对数据进行关联性分析的方法,以及通过可测或易测的变量对未知或不可测变量的状态进行估计的方法。

10.1 线性与趋势性相关分析

对不同类型的变量,用来描述它们之间相互关联的统计量是不同的,关于这一点,在第 7 章中,我们已经有了充分的认识。

事实上,即使对每个研究对象同时观测到的两个指标的成对数据都是连续的,但对其进行相关分析时,在正态分布和非正态分布的前提下,描述两个变量间关联性的相关统计量也是各不相同的。

本节主要讨论两个成对的连续型尺度变量之间的相关分析问题。涉及的方法中,有些同样适用于有序变量间的相关分析。

当两个连续型变量之间不存在确定的函数关系,而又确实存在着某种数量上的关系时,称这类非确定性的关系为相关关系,简称这类变量间存在“相关”。

描述上述两个变量间相关密切程度的量称简单相关系数。两个变量总体间的简单相

关系数用 ρ 表示。两个变量样本间的简单相关系数用符号 r 表示, 显然, 简单相关系数应满足: $0 \leq |r| \leq 1$ 。

之所以称为简单相关, 是由于讨论的变量只有两个, 这是众多变量间相关关系中最简单的一种。当考察的变量为多个时, 在固定其他变量的情况下, 讨论两个变量的相关关系, 显然有别于它, 此时, 称偏相关。而讨论一个变量与其他多个变量的线性组合间的相关关系时, 称复相关。

关于相关, 一般分为线性相关和非线性相关两类。

(1) 线性相关: 凡两个变量在直角坐标系中的散点图围绕一条直线四周进行分布的, 则称这两个变量呈线性相关。

(2) 非线性相关: 凡两个变量在直角坐标系中的散点图围绕一条曲线四周进行分布的, 则称这两个变量呈非线性相关。

处理上述两种相关的数学方法各不相同, 相差甚远。在本书中, 只讨论线性相关, 以下简称相关。

现在, 对于相关的定义还有另一层含义, 当两个变量间存在某种趋势时, 也称两个变量间存在相关。如, X 增加时, Y 增大, 称 X 与 Y 间有正相关, 而 X 增加时, Y 减少, 称 X 与 Y 间有负相关。这种表示两个变量之间存在趋势的相关, 我们称它为趋势相关。

初看起来, 线性相关与趋势性相关似乎是一致的, 实则不同。虽然, 当两个变量间存在线性相关时, X 增加时, Y 也增大 (或减少), 两者没有多少区别, 但两个变量之间不存在线性相关时, 并不表示两个变量之间就不存在趋势性相关。

所以描述趋势性的相关, 既不同于线性相关, 也不同于非线性相关。它是相关的另一种形式。

由此可见, 所谓的两个变量之间的不相关, 是指两个变量之间不存在线性关系或趋势性关系, 它们还可能存在非线性关系, 因而, 仅凭线性相关关系或趋势性相关关系的研究是不能推出两个变量之间没有关系的统计结论的。

10.1.1 Pearson (皮尔逊) 相关系数

1. 适用条件

当两个变量 x 与 y 的总体服从或近似服从正态分布时, 变量 x 与 y 间的线性相关系数用 Pearson 相关系数来计算。

2. Pearson 相关系数的计算

Pearson 定义的相关系数的计算公式为

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

其中, \bar{x} 、 \bar{y} 分别是变量 x 、 y 的均值。 x_i 、 y_i 分别是变量 x 、 y 的第 i 个观测值。

由上面的计算公式不难看出, 相关系数数值的范围在 $-1 \sim 1$ 之间, 相关系数的绝对值越大表明两变量间的线性相关程度越高, 当 $|r|=1$ 时, 表示观察值在直角坐标系中的散点图都在一条直线上, 称为完全相关。这时的相关关系便成了函数关系。当 $r>0$ 时, 称正相关, 反之为负相关。当 $r=0$ 时, 称两个变量间不呈线性关系, 习惯上说这两个变量不相关, 但这并不意味着这两个变量间无关系。

一般 $|r| \geq 0.70$, 称两个变量高度相关, $0.70 > |r| \geq 0.50$, 称两个变量中度相关, $0.50 > |r| \geq 0.30$, 称两个变量低度相关。

3. Pearson 相关系数的检验

由于据以分析的数据是来自某个总体的样本值, 由此算得的相关系数只是总体相关系数 ρ 的一个估计值。所以, 从同一个总体中抽取的不同样本中会得到不同的样本相关系数, 即样本相关系数之间也存在变异性。因此, 必须进行检验。

检验的原假设为 $H_0: \rho=0$ (总体中两个变量间的线性相关系数为 0)。

用来进行检验的统计量为

$$t = \frac{\sqrt{n-2}}{\sqrt{1-r^2}} r$$

在原假设为真时, $t \sim t(n-2)$ 。

其中, r 是样本相关系数, n 是样本观测量数, $n-2$ 是自由度。在原假设下, 当观测的显著性水平小于 0.05 时, 拒绝原假设, 认为两个变量之间存在线性相关; 否则不拒绝两变量间相关系数为 0 的原假设, 即认为两个变量间不存在线性相关关系。

4. Pearson 相关系数的实例分析

例 10.1 在考察弹簧-质量系统中, 做过这样的—个试验, 将不同质量的重物置于垂直悬挂着的弹簧末端, 记录弹簧对应不同重物质量下被拉长的长度, 见表 10-1。试分析弹簧伸长与重物质量之间是否存在线性相关关系。

表 10-1 弹簧-质量系统

质量	0	50	100	150	200	250	300	350	400	450	500
伸长	0	1.000	1.875	2.750	3.250	4.375	4.875	5.675	6.500	7.250	8.000

【题析】 从数据表象来看, 重物的质量越大, 弹簧的伸长长度也越长, 是否存在线性相关关系, 可在直角坐标系中观察其散点图而知。

1. 制作散点图

具体操作步骤如下：


(1) 在 SPSS 数据编辑窗口中，打开数据文件 data10-01.sav。

(2) 选择简单散点图。

按 Graphs→Legacy Dialogs→Scatter/Dot 顺序打开→Scatter/Dot 对话框，见图 2-78。单击 Simple Scatter 图标，选择 Simple Scatter 图，单击 Define 按钮，打开 Simple Scatterplot 对话框，见图 2-79。

(3) 定义 Y 轴和 X 轴的变量。

将弹簧拉长长度移入 Y Axis 框中，将质量移入到 X Axis 框中。

单击 OK 按钮，在输出窗口中，得到散点图，双击图形，打开图形编辑器，见图 10-1，在工具栏中单击  按钮，在坐标轴中插入拟合斜线，关闭图形编辑器，则在输出窗口中得到添加了拟合直线的散点图，见图 10-2。

从图 10-2 两变量散点图中可见，各点之间基本成一直线，所以弹簧伸长长度与质量之间很有可能存在线性相关关系。

所以，对本类问题，可作线性相关的假设检验。

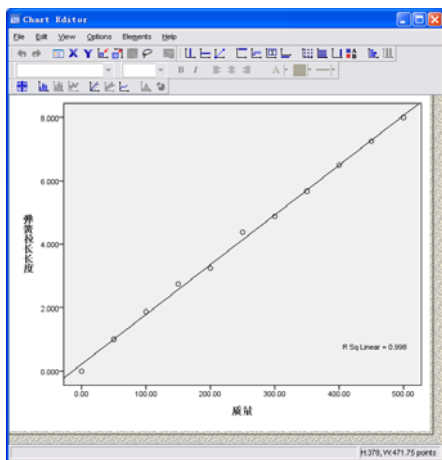


图 10-1 Chart Editor 窗口

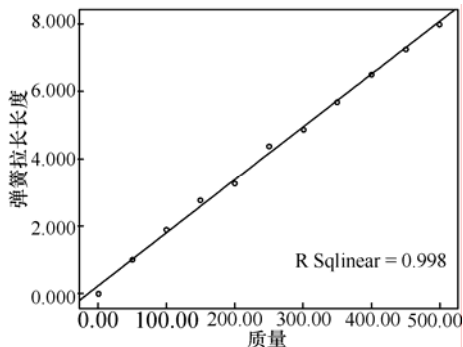


图 10-2 弹簧-质量系统散点图

2. 作数据资料的正态性检验

根据第 2 章 2.4.2 中 3 检查计量的正态性和方差齐性中所述，因为本例中样本含量为 11，属小样本，又由于总体均数未知，需用样本均数来估计，故可用以下步骤来作正态性检验：

(1) 按 Analyze→Descriptive Statistics→Explore 顺序打开 Explore 对话框，见图 2-99。

(2) 选择检验变量

将弹簧拉长长度、质量移入到 Dependent List 下框中。

(3) 选择 Lilliefors 法数据资料正态性检验

单击 Plots 按钮，展开 Plots 对话框，见图 2-104。选择 Normality plots with tests 选项，要求在输出窗输出图形与检验结果。单击 Continue 按钮返回图 2-99。

(4) 单击 OK 按钮，在输出窗中得到表 10-2 的统计算结果。

(5) 结论

表 10-2 原始数据的正态性检验

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
质量	.090	11	.200 [*]	.968	11	.870
弹簧拉长长度	.090	11	.200 [*]	.973	11	.913

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

因为，在总体服从正态分布的原假设下，观测的显著性水平为 0.200 以上，大于 0.05，所以现有证据不支持拒绝弹簧拉长长度、质量变量服从正态分布的假设。

3. 计算简单相关系数及其进行相关系数显著性检验

继续在 SPSS 中进行如下操作步骤：

(1) 选择分析变量

按 Analyze→Correlate→Bivariate(二元变量)顺序，打开二元变量相关分析对话框，见图 10-3。选中对话框左面的变量表中的弹簧拉长长度、质量变量，将其移入到 Variables 矩形框中。

(2) 选择相关系数计算公式

由于本例变量不拒绝服从正态分布，所以，在 Correlation Coefficients 下，选择 Pearson。

(3) 确定显著性检验方法

在 Test of Significance 下，选择 Two-tailed（双侧 t 检验）。

本例从散点图中已知相关的方向为正相关，所以也可以选择 One tailed（单侧 t 检验）。

(4) 标识显著性

选择 Flag significant Correlations 选项，要求在输出结果中，相关系数右上方使用 “*” 表示显著性水平为 5%，用 “**” 表示其显著性水平为 1%。

(5) 选择 Options

单击 Options 按钮，打开 Options 对话框，见图 10-4。

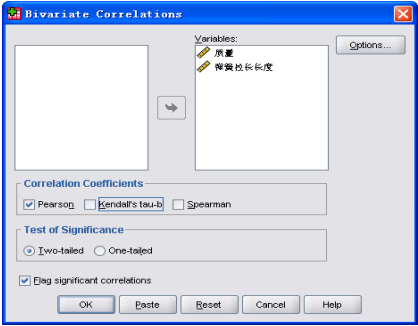


图 10-3 二元变量相关分析对话框

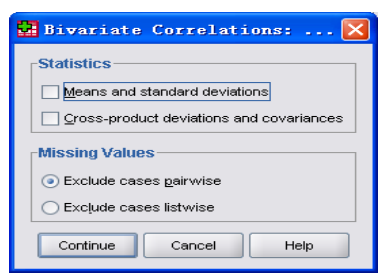


图 10-4 输出选择项对话框

Statistics (统计量) 选择项中有两个有关统计量的选择项。它只在对话框中选择了 Pearson 相关分析方法时才可用。如果选择 Means and standard deviations 选项, 则要求计算并输出均值与标准差。如果选择 Cross-product deviations and covariances 选项, 则要求计算并输出叉积离差阵和协方差阵。

本例目的只是计算相关系数, 故不作选择。

在 Missing Values 栏中有两个关于缺失值处理方法的选择项: 如果选择 Exclude cases pairwise 选项, 则仅剔除正在参与计算的两个变量值是缺失值的观测量。这样计算出的相关系数矩阵, 相关系数是根据不同数量的观测量计算出来的。选择此项, 可以最大限度地使用取得的观测数据。如果选择 Exclude cases listwise 选项, 则剔除在主对话框 Variables 矩形框中列出的变量带有缺失值的所有观测量。这样计算出的相关系数矩阵, 每个相关系数都是依据相同数量的观测量计算出来的。

在本例中, 这两个选择项都可选, 故直接采用系统默认的第一项。

(6) 单击 OK 按钮运行, 在输出窗口中得到计算结果, 见表 10-3。

从表 10-3 可见, 质量与弹簧拉长长度之间的相关系数 $r=0.999$, 样本含量 $N=11$, 在 $H_0:\rho=0$ 原假设下, 观测的显著性水平为 0.0000, 小于 0.05, 故拒绝原假设, 而认为质量与弹簧拉长长度之间的线性相关关系在统计上有极显著性意义。

表 10-3 相关系数计算结果

Correlations			
质量		质量	弹簧拉长长度
	Pearson Correlation	1	.999**
	Sig. (2-tailed)		.000
N		11	11
弹簧拉长长度		质量	弹簧拉长长度
	Pearson Correlation	.999**	1
	Sig. (2-tailed)	.000	
N		11	11

** . Correlation is significant at the 0.01 level (2-tailed).

10.1.2 Spearman (斯皮尔曼) 秩相关

秩相关的方法由 C.Spearman 于 1904 年提出。秩相关系数可以用来描述两个变量有没有同时上升 (下降), 或一个上升、一个下降的趋势。

1. Spearman 秩相关的适用条件

当成对的两个数值型变量的总体不服从正态分布或总体分布未知时, 由于不符合 Pearson 相关的条件, 所以不能用积差相关系数来刻画相关性, 此时, 可先将两个变量转换成有序变量, 计算各自的秩, 再用 Spearman 秩相关系数进行计算。它属非参数统计方法。

同样, 两个本身就是成对有序变量之间的相关关系也可用 Spearman 秩相关进行分析。

2. Spearman 秩相关系数的计算

设有成对数据

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

记 x_i 在 $\{x_1, x_2, \dots, x_n\}$ 中的秩为 R_i , y_i 在 $\{y_1, y_2, \dots, y_n\}$ 中的秩为 Q_i , $i=1, 2, \dots, n$ 。Spearman 定义的秩相关系数的计算公式为

$$r_s = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (S_i - \bar{S})^2}} = 1 - \frac{6 \sum_{i=1}^n (R_i - S_i)^2}{n(n^2 - 1)}$$

它相当于用 R_i 和 Q_i 替代原始的 x_i 和 y_i 后的 Pearson 相关系数。

在有结时, 即有相等的观察值时, 秩取平均值, 则 Spearman 定义的秩相关系数的计算公式为

$$r_s = \frac{12 \sum_{i=1}^n a(R_i) b(S_i) - 3n(n+1)^2}{\sqrt{n(n^2 - 1) - \sum_{t=1}^{g_x} (\tau_{x,t}^3 - T_{x,t})} \sqrt{n(n^2 - 1) - \sum_{t=1}^{g_y} (\tau_{y,t}^3 - T_{y,t})}}$$

其中, $a(r)$, $r=1, 2, \dots, n$ 为 x_i 在 $\{x_1, x_2, \dots, x_n\}$ 中秩 R_i 的计分函数, $b(r)$, $r=1, 2, \dots, n$ 为 y_i 在 $\{y_1, y_2, \dots, y_n\}$ 中秩 S_i 的计分函数。在结的长度为 1 时, $a(R_i) = R_i$, $b(S_i) = S_i$, 而结的长度超过 1 时, $a(R_i)$ 、 $b(S_i)$ 等于秩的平均。 g_x 为样本数据 $\{x_1, x_2, \dots, x_n\}$ 中的结的个数, $\tau_{x,t}$ 是样本数据 $\{x_1, x_2, \dots, x_n\}$ 中第 t 个结的长度, $t=1, 2, \dots, g_x$; 而 g_y 为样本数据 $\{y_1, y_2, \dots, y_n\}$ 中的结的个数, $\tau_{y,t}$ 是样本数据 $\{y_1, y_2, \dots, y_n\}$ 中第 t 个结的长度, $t=1, 2, \dots, g_y$ 。

3. Spearman 秩相关系数的检验

在 X 与 Y 相互独立的原假设为真时, 秩相系数 r_s 服从对称分布, 对称中心为原点 0。它的期望和方差分别为

$$E(r_s) = 0, \quad D(r_s) = 1/(n-1)$$

可以证明: 在 $n \rightarrow \infty$ 时, 秩相关系数 r_s 有渐近正态性

$$\sqrt{n-1} r_s \xrightarrow{L} N(0, 1), \quad n \rightarrow \infty$$

由此, 在大样本情况下, 可用正态分布原理对其进行检验。

小样本时, Spearman 秩相关系数的检验方法, 同 Pearson 相关系数的检验。

4. Spearman 秩相关系数的实例分析

例 10.2 从某校高二男生中随机抽测 10 名学生进行引体向上 x 和双杠双臂屈伸 y 的测试, 结果见表 10-4。试问引体向上和双杠双臂屈伸这两个变量之间是否存在相关?

表 10-4 引体向上和双杠双臂屈伸记录表

引体向上	9	12	13	11	16	28	13	14	16	5
双杠双臂屈伸	16	13	17	13	18	30	16	17	15	9

1. 数据资料正态性检验

在 SPSS 数据编辑窗口中, 将上述数据简称为数据文件, 见 data10-02.sav。

按例 10.1 中的数据资料正态性检验的方法, 得如下计算结果, 见表 10-5。由于双杠双臂屈伸服从正态分布的原假设成立的概率为 $0.021 < 0.05$, 所以, 双杠双臂屈伸不服从正态分布。故要求引体向上和双杠双臂屈伸之间的相关关系不能采用 Pearson 相关。

表 10-5 原始数据正态性检验

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
引体向上	.251	10	.075	.872	10	.107
双杠屈伸	.285	10	.021	.816	10	.022

a. Lilliefors Significance Correction

2. 计算简单相关系数及其进行相关系数显著性检验

继续在 SPSS 中进行如下操作步骤:

(1) 选择分析变量

按 Analyze→Correlate→Bivariate (二元变量) 顺序, 打开二元变量相关分析对话框, 见图 10-2。选中对话框左面的变量表中的 *引体向上*、*双杠屈伸* 变量, 将其移入到 Variables 矩形框中。

(2) 选择相关系数计算公式

由于本例变量不服从正态分布, 所以, 在 Correlation Coefficients 下, 选择 Spearman。

(3) 确定显著性检验方法

在 Test of Significance 下, 选择 Two-tailed (双侧 t 检验)。

(4) 标识显著性

选择 Flag significant Correlations 选项, 要求在输出结果中, 相关系数右上方使用 “*” 表示显著性水平为 5%, 用 “**” 表示其显著性水平为 1%。

(5) 单击 OK 按钮运行, 则在输出窗口中得到如下计算结果, 见表 10-6。

表 10-6 相关系数计算结果

Correlations			引体向上	双杠屈伸
Spearman's rho	引体向上	Correlation Coefficient	1.000	.757*
		Sig. (2-tailed)	.	.011
		N	10	10
	双杠屈伸	Correlation Coefficient	.757*	1.000
		Sig. (2-tailed)	.011	.
		N	10	10

*. Correlation is significant at the 0.05 level (2-tailed).

从表 10-6 可见,引体向上与双杠双臂屈伸之间的 Spearman 相关系数为 0.757,在 $\rho=0$ 的原假设下,观测的显著性水平为 $0.011<0.05$,所以,拒绝原假设,而认为引体向上与双杠双臂屈伸之间有正相关,即引体向上次数越多,相应的双杠双臂屈伸的次数一般也越多,反之亦然。

10.1.3 Kendall's tau-b (肯德尔 τ -b) 相关系数

Kendall's tau-b 相关系数是 Kendall 在 1938 年提出的,用来检验二元连续型随机变量之间的相关性。它可以用来描述两个变量有没有同时上升(下降),或对应于一个上升、另一个下降的趋势。

1. Kendall's tau-b 相关系数的适用条件

当成对的两个数值型变量的总体不服从正态分布或总体分布未知时,此时,不能用 Pearson 相关来刻画两个变量的相关性,而应改用 Spearman 秩相关系数,同样也可用 Kendall's tau-b 相关系数来描述两个变量的相关性。它也属于非参数统计方法。

同样,两个本身就是成对有序变量之间的相关关系也可用 Kendall's tau-b 相关进行分析。

2. Kendall's tau-b 相关系数的计算

设有成对数据

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

是独立同分布样本,其总体为 (X,Y) 。当 X 增加时, Y 有增大的趋势,称正相关。这也就是在 $x_2 > x_1$ 时, y_2 有大于 y_1 的趋势,换言之,“ $(x_2 - x_1)(y_2 - y_1) > 0$ ”发生的可能性超过“ $(x_2 - x_1)(y_2 - y_1) < 0$ ”发生的可能性。

$$\text{令 } \theta = P((x_2 - x_1)(y_2 - y_1) > 0) - P((x_2 - x_1)(y_2 - y_1) < 0)$$

则在 X 与 Y 存在正相关时, θ 的值大于0。同理,在 X 与 Y 存在负相关时, θ 的值小于0;而在 X 与 Y 不存在相关时, θ 的值接近于0。所以, θ 是表示 X 与 Y 相关性的一个总体参数。

Kendall's tau-b 是用来估计 θ 的统计量的, 它被定义为

$$\tau = \frac{2}{n(n-1)} z, \quad z = \sum_{1 \leq i < j \leq n} \text{sign}((x_j - x_i)(y_j - y_i))$$

其中, 符号函数 sign 的定义如下:

$$\text{sign}(t) = \begin{cases} 1, & t > 0 \\ 0, & t = 0 \\ -1, & t < 0 \end{cases}$$

由此可见, τ 相关系数仅与 x_1, x_2, \dots, x_n 之间, 以及 y_1, y_2, \dots, y_n 之间的大小顺序有关, 而与它们的数值差值的大小无关。

因此, 设在 x_1, x_2, \dots, x_n , 以及在 y_1, y_2, \dots, y_n 中都没有重复的观察数据, x_i 在 $\{x_1, x_2, \dots, x_n\}$ 中的秩为 R_i , $R_i = 1, 2, \dots, n$, y_i 在 $\{y_1, y_2, \dots, y_n\}$ 中的秩为 Q_i , $Q_i = 1, 2, \dots, n$ 。

则有: $z = \sum_{1 \leq i < j \leq n} \text{sign}((R_j - R_i)(Q_j - Q_i))$, 在 X 与 Y 相互独立的原假设下, τ 相关系数和 $\tilde{\tau}$ 同分布, 其中

$$\tilde{\tau} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \text{sign}(R_j - R_i) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \text{sign}(x_j - x_i)$$

当有重复的观测值时, 即有结时, 秩取平均。此时, 要对 z 表达式进行修改

$$\begin{aligned} z &= \sum_{1 \leq i < j \leq n} \text{sign}((a(R_j) - a(R_i))(b(Q_j) - b(Q_i))) \\ &= \sum_{1 \leq i < j \leq n} \text{sign}(a(R_j) - a(R_i)) \text{sign}(b(Q_j) - b(Q_i)) \end{aligned}$$

其中, $a(r)$ 、 $b(r)$ 为计分函数, 其含义的解释参见在 Spearman 秩相关系数的计算中所提到的。

根据不等式 $(\sum c_i d_i)^2 \leq \sum c_i^2 \sum d_i^2$, 可以得到

$$\begin{aligned} z^2 &\leq \sum_{1 \leq i < j \leq n} [\text{sign}(a(R_j) - a(R_i))]^2 \sum_{1 \leq i < j \leq n} [\text{sign}(b(Q_j) - b(Q_i))]^2 \\ &= \left(\frac{n(n-1)}{2} - \frac{\sum_{t=1}^{g_x} \tau_{x,t} (\tau_{x,t} - 1)}{2} \right) \left(\frac{n(n-1)}{2} - \frac{\sum_{t=1}^{g_y} \tau_{y,t} (\tau_{y,t} - 1)}{2} \right) \end{aligned}$$

式中的含义同 Spearman 秩相关系数的计算中所提到的。

由此, 在有结时, τ 相关系数的定义为

$$\tau = \frac{z}{\sqrt{\frac{n(n-1)}{2} - \frac{\sum_{t=1}^{g_x} \tau_{x,t}(\tau_{x,t} - 1)}{2}} \sqrt{\frac{n(n-1)}{2} - \frac{\sum_{t=1}^{g_y} \tau_{y,t}(\tau_{y,t} - 1)}{2}}} \\ = \frac{\sum_{1 \leq i < j \leq n} \text{sign}(a(R_j) - a(R_i)) \text{sign}(b(Q_j) - b(Q_i))}{\sqrt{\frac{n(n-1)}{2} - \frac{\sum_{t=1}^{g_x} \tau_{x,t}(\tau_{x,t} - 1)}{2}} \sqrt{\frac{n(n-1)}{2} - \frac{\sum_{t=1}^{g_y} \tau_{y,t}(\tau_{y,t} - 1)}{2}}}$$

3. Kendall's tau-b 相关系数的检验

可以证明，在原假设为真时，即 X 与 Y 相互独立时， τ 相关系数有渐近正态性

$$3 \sqrt{\frac{n(n-1)}{2(2n+5)}} \tau \xrightarrow{L} N(0,1), \quad n \rightarrow \infty$$

因此，在样本含量 n 较大时，可以用渐近正态性计算得到原假设成立的概率值。

4. Kendall's tau-b 相关系数的实例分析

例 10.3 从 2007 年环海南国际公路多日赛的比赛成绩册中，随机抽取平路赛第二段与山路赛第五段，并从山路赛第五段中随机抽取 14 名运动员，得到两赛段的运动员的名次见表 10-7。试问平路赛段成绩好的运动员是否也意味着他的山路赛段成绩也好？

表 10-7 14 名运动员的平路和山路赛段名次表

运动员号码	101	171	174	111	26	42	21	62	92	32	115	3	4	6
第二段名次	3	5	2	58	55	6	70	1	57	33	53	79	23	78
第五段名次	70	6	23	5	20	10	18	56	21	12	34	9	2	55

【题析】 本例想要观察的是平路赛段的名次与山路赛段的名次之间是否正相关、负相关，还是不相关。

由于本例的两个变量都是有序变量，所以可以用 Kendall's tau-b 或 Spearman 相关来计算，但显然不适合用 Pearson 相关。

因此，在 SPSS 中可用如下步骤来分析处理：

1. 在 SPSS 中建立数据文件，见 data10-03.sav。
2. 选择分析变量

按 Analyze→Correlate→Bivariate（二元变量）顺序，打开二元变量相关分析对话框，见图 10-2。选中对话框左面的变量表中的第二段平路名次、第五段山路名次变量，将其移入到 Variables：矩形框中。

3. 选择相关系数计算公式

由于本例变量是典型的分类变量, 所以, 在 Correlation Coefficients 下, 选择 Kendall's tau-b。

4. 确定显著性检验方法

在 Test of Significance 下, 选择 Two-tailed (双侧 t 检验)。

5. 标识显著性

选择 Flag significant Correlations 选项, 要求在输出结果中, 相关系数右上方使用 “*” 表示显著性水平为 5%, 用 “**” 表示其显著性水平为 1%。

6. 单击 OK 按钮运行, 则在输出窗口中得到如下结果, 见表 10-8。

因为第二段平路名次、第五段山路名次的 Kendall's tau-b 相关系数为-0.143, 在 $\rho = 0$ 的原假设下, 观测的显著性水平为 $0.477 > 0.05$, 所以, 现有证据不支持拒绝原假设, 而认为第二段平路名次、第五段山路名次之间不存在相关关系, 即运动员平路赛段名次的好坏与山路赛段的名次好坏不存在对应关系。

表 10-8 Kendall's tau-b 计算结果

Correlations			第二段平路名次	第五爬山赛段名次
Kendall's tau_b	第二段平路名次	Correlation Coefficient	1.000	-.143
		Sig. (2-tailed)	.	.477
		N	14	14
	第五爬山赛段名次	Correlation Coefficient	-.143	1.000
		Sig. (2-tailed)	.477	.
		N	14	14

10.2 偏相关分析

10.2.1 偏相关的概念

1. 偏相关分析

简单相关系数, 只是反映两个变量间线性关系密切的程度, 并不一定能真正揭示两个变量之间的内在关系。例如身高、体重与肺活量之间的关系, 从简单相关的角度可以得到肺活量与身高和体重均存在较强的线性关系的统计结论。但如果我们控制体重, 再分析身高和肺活量之间的关系, 就会得到与简单相关相反的结论。而后者更符合实际, 因为肺活量大小与胸围大小之间存在正相关, 当体重一定时, 胸围小者身高高, 所以身高与肺活量为负相关是符合实际的。

这种控制一个变量再去讨论其他变量之间线性关系密切程度的方法称为偏相关分析。同样, 称此时描述其他两个变量之间线性关系密切程度的量为偏相关系数。

2. 偏相关系数的计算

设控制变量 z , 则变量 x 、 y 之间的偏相关系数的计算公式为

$$r_{xy,z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}$$

式中 $r_{xy,z}$ 是控制了 z 的条件下, x 、 y 之间的偏相关系数。 r_{xy} 是变量 x 、 y 间的简单相关系数。 r_{xz} 、 r_{yz} 分别是变量 x 、 z 间的和变量 y 、 z 间的简单相关系数。

当控制了二个变量 z_1 、 z_2 时, 变量 x 、 y 之间的偏相关系数的计算公式为

$$r_{xy,z_1 z_2} = \frac{r_{xy,z_1} - r_{xz_2,z_1} r_{yz_2,z_1}}{\sqrt{(1-r_{xz_2,z_1}^2)(1-r_{yz_2,z_1}^2)}}$$

式中含义同上式。

3. 偏相关系数的检验

H_0 : 总体中两个变量间的偏相关系数为 0。

检验的统计量为: $t = \frac{\sqrt{n-k-2}}{\sqrt{1-r^2}} r$

当原假设为真时, $t \sim t(n-k-2)$

其中, r 是相应的偏相关系数, n 是样本含量, k 是控制变量的数目, $n-k-2$ 是自由度。

在原假设下, 当观测的显著性水平小于 0.05 时, 拒绝原假设, 否则, 没有充分的证据拒绝两变量总体间偏相关系数为 0 的原假设。

10.2.2 偏相关实例分析

例 10.4 随机测试了北京市海淀区某初中 13 岁男孩的身高、体重、肺活量和胸围的数据资料。存放在数据文件 data10-04.sav 中。试进行相关与偏相关分析。

在 SPSS 中具体操作步骤如下:

1. 打开 data10-04.sav。

2. 按例 10.1 中的方法对身高、体重、肺活量、胸围变量进行正态性检验。结果见表 10-9。

由于 4 个变量服从正态分布的原假设的概率分为为: 0.200、0.200、0.121、0.200, 均大于 0.05, 所以不能拒绝它们服从正态分布的原假设。

3. 按 Analyze→Correlation→Bivariate (二元变量) 顺序, 打开二元变量相关分析对话框, 见图 10-2。选中对话框左面的变量表中的身高、体重、肺活量、胸围变量, 将其移入到 Variables 矩形框中。

4. 选择相关系数计算公式

由于本例变量都服从正态分布, 所以, 在 Correlation Coefficients 下, 选择 Pearson。

5. 其他保持系统默认选项, 单击 OK 按钮, 在输出窗口中得到表 10-10 的输出结果。

表 10-9 身高等 4 个变量的正态性检验结果

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
身高	.081	29	.200 [*]	.977	29	.762
体重	.099	29	.200 [*]	.961	29	.357
肺活量	.145	29	.121	.942	29	.111
胸围	.095	29	.200 [*]	.976	29	.732

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

表 10-10 相关系数

Correlations					
	身高	体重	肺活量	胸围	
身高	1	.741**	.600**	.814**	
	Pearson Correlation				
	Sig. (2-tailed)	.000	.001	.000	
	N	29	29	29	
体重	.741**	1	.751**	.834**	
	Pearson Correlation				
	Sig. (2-tailed)	.000	.000	.000	
	N	29	29	29	
肺活量	.600**	.751**	1	.805**	
	Pearson Correlation				
	Sig. (2-tailed)	.001	.000	.000	
	N	29	29	29	
胸围	.814**	.834**	.805**	1	
	Pearson Correlation				
	Sig. (2-tailed)	.000	.000	.000	
	N	29	29	29	

**. Correlation is significant at the 0.01 level (2-tailed).

6. 指定分析变量和控制变量

按 Analyze→Correlation→Partial 顺序, 打开偏相关分析对话框。见图 10-5。

将身高、肺活量、胸围变量移到 Variables 矩形框中。将体重变量移到 Controlling for 矩形框中。设定体重变量为控制变量。

7. 单击 OK 按钮, 在输出窗中, 得到如表 10-11 所示的输出结果。

表 10-11 偏相关系数

Correlations				
Control Variables	身高	肺活量	胸围	
体重 身高	Correlation	1.000	.098	.527
	Significance (2-tailed)		.619	.004
	df	0	26	26
肺活量	Correlation	.098	1.000	.490
	Significance (2-tailed)	.619		.008
	df	26	0	26
胸围	Correlation	.527	.490	1.000
	Significance (2-tailed)	.004	.008	
	df	26	26	0

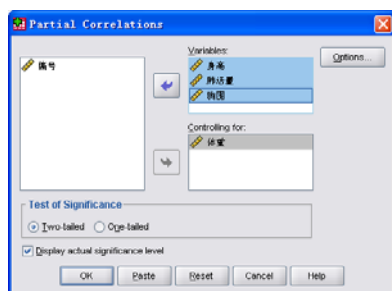


图 10-5 偏相关对话框

8. 分析结果解释与结论

从表 10-10 中可见, 身高、体重、肺活量与胸围两两之间都存在正相关, 且相关都具有显著性意义。

从表 10-11 中看见, 当控制体重变量后, 身高与肺活量之间不存在相关关系 ($P=0.619>0.05$), 而胸围与肺活量之间存在正相关 ($P=0.008<0.01$), 相关有统计上的显著性意义。这说明, 当体重一定时, 胸围越大, 肺活量也越大。

10.3 距离分析

10.3.1 距离分析概述

10.3.1.1 概念

距离分析是通过计算成对样品之间或成对变量之间的广义距离, 对它们间的相似或

不相似（距离）的程度进行测度，以考察其相互接近程度的一种统计方法。

距离分析的结果可以用于因子分析、聚类分析等，有助于分析复杂的数据集。

10.3.1.2 分类

1. 根据分析对象的不同，距离分析分为两种：
 - 样品间距离分析：样品和样品之间的距离分析。
 - 变量间距离分析：变量和变量之间的距离分析。
2. 根据测度的统计量的不同，距离分析分为两种：
 - 不相似性测度：通过计算样品之间或变量之间的距离来表示。
 - 相似性测度：通过计算 Pearson 相关系数或 Cosine 相似系数来表示。

10.3.1.3 适用于不同类型变量的不相似测度

1. 对连续型变量进行距离分析时，常用的不相似测度的统计量有以下几种：

(1) 欧氏距离 (Euclidean Distance)

计算公式为

$$EUCLID = \sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

式中， k 每个样品中有 k 变量（测试指标）； x_i 表示第一个样品在第 i 测试指标上的取值， y_i 表示第二个样品在第 i 个测试指标上的取值。

(2) 欧氏距离平方 (Squared Euclidean Distance)

计算公式为

$$SEUCLID = \sum_{i=1}^k (x_i - y_i)^2$$

式中符号的含义同欧氏距离中的说明。

(3) 切贝谢夫距离 (Chebychev Distance)

计算公式为

$$CHEBYCHEV(x, y) = \max |x_i - y_i|$$

式中 $\max()$ 为取最大值函数，其他符号的含义同欧氏距离中的说明。

(4) 布洛克距离 (Block Distance)

计算公式为

$$BLOCK(x, y) = \sum_{i=1}^k |x_i - y_i|$$

式中符号的含义同欧氏距离中的说明。

(5) 明可斯基距离 (Minkowski Distance)

计算公式为

$$MINKOWSKI(x, y) = \sqrt[p]{\sum_{i=1}^k |x_i - y_i|^p}$$

式中, p 是任意指定的次方, 其他符号的含义同欧氏距离中的说明。

(6) 用户自定义距离 (Customized Distance)

计算公式为

$$Customized(x, y) = \sqrt[r]{\sum_{i=1}^k |x_i - y_i|^p}$$

式中, r 、 p 是任意指定的次方, 其他符号的含义同欧氏距离中的说明。

2. 对离散型 (有序或名义) 变量进行距离分析时, 常用的不相似测度的统计量有如下几种:

(1) 卡方测度 (Chi-square measure)

计算公式为

$$CHISQ(x, y) = \sqrt{\frac{\sum_{i=1}^k (x_i - E(x_i))^2}{E(x_i)} + \frac{\sum_{i=1}^k (y_i - E(y_i))^2}{E(y_i)}}$$

式中, $E(x_i)$ 和 $E(y_i)$ 分别是 x_i 和 y_i 的理论期望频数, 其他符号的含义同欧氏距离中的说明。

(2) 斐方 (ϕ^2) 测度 (Phi-square measure)

计算公式为

$$PHISQ(x, y) = \sqrt{\frac{\sum_{i=1}^k (x_i - E(x_i))^2}{E(x_i)} + \frac{\sum_{i=1}^k (y_i - E(y_i))^2}{E(y_i)}} = \frac{CHISQ(x, y)}{\sqrt{n}}$$

式中符号的含义同卡方测度。

表 10-12 整理成的四格表

	第二特性	
	发生	不发生
第一特性		
发生	a	b
不发生	c	d

3. 对二值变量所进行的距离分析是基于整理后所得的表 10-12 所示的四格表的基础上的。常用的不相似测度的统计量有如下几种:

(1) 二值欧氏距离

计算公式为

$$\text{二值欧氏距离} = \sqrt{b + c}$$

式中, b 、 c 表示事件在了一项中发生, 但在另一项中不发生的四格表中对角线单元。其最小值为 0, 最大值不限。

(2) 二值欧氏距离平方

计算公式为

$$\text{二值欧氏距离平方} = b + c$$

式中符号含义同二值欧氏距离中的说明。它计算的是不匹配事件的数量。其最小值为 0，最大值不限。

(3) 不对称指数 (Size difference)

计算公式为

$$SIZE(x, y) = \frac{(b - c)^2}{n^2}$$

式中， b 、 c 的含义同二值欧氏距离中的说明， n 为总观察次数。其值在 0~1 之间。

(4) Pattern difference

计算公式为

$$PATTERN(x, y) = \frac{bc}{n^2}$$

式中， b 、 c 表示事件在一项中发生，但在另一项中不发生的四格表中对角线单元， n 为总观察次数。其最小值为 0，最大值为 1。

(5) 方差 (Variance)

计算公式为

$$VARIANCE(x, y) = \frac{b + c}{4n}$$

式中， b 、 c 表示事件在一项中发生，但在另一项中不发生的四格表中对角线单元， n 为总观察次数。其最小值为 0，最大值为 1。

(6) Shape

计算公式为

$$BSHAPE(x, y) = \frac{n(b + c) - (b - c)^2}{n^2}$$

式中， b 、 c 表示事件在一项中发生，但在另一项中不发生的四格表中对角线单元， n 为总观察次数。其取值范围无上下限。

(7) Lance and Williams

计算公式为

$$LANCEANDWILLIAMS(x, y) = \frac{b + c}{2a + b + c}$$

式中， a 表示事件在两项中都发生的相应单元， b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元。其最小值为 0，最大值为 1。

10.3.1.4 适用于不同类型变量的相似测度

1. 对连续型变量进行距离分析时, 常用的相似测度的统计量有以下几种:

(1) Pearson 相关

计算公式及含义同 10.1.1。

(2) Cosine 相似系数 (Cosine similarity measure)

计算公式为

$$COSINE(x, y) = \frac{\sum_{i=1}^k (x_i y_i)^2}{\sqrt{(\sum_{i=1}^k x_i^2)(\sum_{i=1}^k y_i^2)}}$$

式中, k 表示每个样品中有 k 个变量 (测试指标); x_i 表示第一个样品在第 i 个测试指标上的取值, y_i 表示第二个样品在第 i 个测试指标上的取值。计算值向量间的余弦, 值范围是 $-1 \sim 1$, 用 0 值表明两向量正交 (相互垂直)。

2. 对于二值变量进行距离分析时, 常用的相似测度的统计量有如下几种:

(1) Russel and Rao

计算公式为

$$RR(x, y) = \frac{a}{a + b + c + d}$$

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在两项中发生而在另一项中不发生的对角线单元, d 表示事件在两项中都不发生的相应单元。其最小值为 0, 最大值为 1。它是内积 (点积) 的二元形式。对匹配与不匹配都给予相等的权重。

(2) Simple matching

计算公式为

$$SM(x, y) = \frac{a + d}{n}$$

式中, a 表示事件在两项中都发生的相应单元, d 表示事件在两项中都不发生的相应单元, n 为总观察次数。其取值范围为 $0 \sim 1$ 。它是匹配数对值的总数的比值。它给匹配与不匹配以相同的权重。

(3) Jaccard

计算公式为

$$JACCARD(x, y) = \frac{a}{a + b + c}$$

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在两项中发生而在另一项中不发生的对角线单元。其最小值为 0, 最大值为 1。它是一个不考虑联合缺席 (d)

的指数。它给匹配与不匹配以相等的权重，同相似比类似。

(4) Dice

计算公式为

$$DICE(x, y) = \frac{2a}{2a + b + c}$$

式中， a 表示事件在两项中都发生的相应单元， b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元。其最小值为 0，最大值为 1。该指数中剔除了联合不发生，给匹配双倍权重。类似于 Czekanowski 或 Sorensen 测度。

(5) Rogers and Tanimoto

计算公式为

$$RT(x, y) = \frac{a + d}{a + d + 2(b + c)}$$

式中， a 表示事件在两项中都发生的相应单元， b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元， d 表示事件在两项中都不发生的相应单元。其最小值为 0，最大值为 1。它是一个给不匹配 (b 、 c) 双倍权重的指数。

(6) Sokal and Sneath 1

计算公式为

$$SS1(x, y) = \frac{2(a + d)}{2(a + d) + b + c}$$

式中， a 表示事件在两项中都发生的相应单元， b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元， d 表示事件在两项中都不发生的相应单元。其最小值为 0，最大值为 1。它是给匹配以双倍权重的一种指数。

(7) Sokal and Sneath 2

计算公式为

$$SS2(x, y) = \frac{a}{a + 2(b + c)}$$

式中， a 表示事件在两项中都发生的相应单元， b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元。其最小值为 0，最大值为 1。它是给不匹配以双倍权重的一种指数，且不考虑联合缺席的情况。

(8) Sokal and Sneath 3

计算公式为

$$SS3(x, y) = \frac{a + d}{b + c}$$

式中， a 表示事件在两项中都发生的相应单元， b 和 c 表示事件在一项中发生而在另

一项中不发生的对角线单元, d 表示事件在两项中都不发生的相应单元。其最小值为 0, 最大值不限。它是匹配与不匹配的比。

(9) Kulczynski 1

计算公式为

$$K1(x, y) = \frac{a}{b + c}$$

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元。其最小值为 0, 最大值不限。它是联合出现与非匹配数的比。

(10) Kulczynski 2

计算公式为

$$K2(x, y) = \frac{a/(a+b) + a/(a+c)}{2}$$

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元。该指数范围为 0~1。它根据某特性在一项中出现的条件概率给出在其他项中出现的概率。计算该指数时, 每一项作为其他项的预测值时, 各值取其平均数。

(11) Sokal and Sneath 4

计算公式为

$$SS4(x, y) = \frac{a/(a+b) + a/(a+c) + d/(b+d) + d/(c+d)}{4}$$

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元, d 表示事件在两项中都不发生的相应单元。该指数范围为 0~1。它是同一匹配状态(某特性出现或不出现)在另一项出现或不出现时的条件概率。计算该指数时, 每一项作为其他项的预测值时, 各项值取其平均数。

(12) Hamann

计算公式为

$$HAMANN(x, y) = \frac{(a+d) - (b+c)}{n}$$

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元, d 表示事件在两项中都不发生的相应单元。该指数范围为 -1~1。

(13) Lambda

计算公式为

$$LAMBDA(x, y) = \frac{t_1 - t_2}{2n - t_2}$$

其中 $t_1 = \text{Max}(a,b) + \text{Max}(c,d) + \text{Max}(a,c) + \text{Max}(b,d)$, $t_2 = \text{Max}(a+c,b+d) + \text{Max}(a+d,c+d)$ 。

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元, d 表示事件在两项中都不发生的相应单元。该指数范围为 0~1。它是 Goodman and Kruskal 中的 λ 。当预测方向同等重要时该系数估计的是使用一项预测另一项的误差降低的比例。

(14) Anderberg'D

计算公式为

$$D(x, y) = \frac{t_1 - t_2}{2n}$$

其中 t_1 、 t_2 定义与 (13) Lambda 中的定义相同。该指数类似于 λ , 它取决于用一项预测另一项 (在两个方向上进行预测) 的误差降低的实际数值。其值范围为 0~1。

(15) Yule's Y

计算公式为

$$Y(x, y) = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元, d 表示事件在两项中都不发生的相应单元。它的取值范围为 -1~1。它是 2×2 表交叉比函数, 且独立于边际总和。

(16) Yule's Q

计算公式为

$$Q(x, y) = \frac{ad - bc}{ad + bc}$$

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元, d 表示事件在两项中都不发生的相应单元。它的取值范围为 -1~1。它是 Goodman 和 Kruskal 中 γ (gamma) 的特殊事件, 是 2×2 表交叉比函数, 且独立于边际总和。

(17) Ochiai

计算公式为

$$OCHIAI(x, y) = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在一项中发生而在另一项中不发生的对角线单元。其取值范围为 0~1。它是余弦相似性测度的二元形式。

(18) Sokal and Sneath 5

计算公式为

$$SS5(x, y) = \frac{ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在两项中发生而在另一项中不发生的对角线单元, d 表示事件在两项中都不发生的相应单元。它的取值范围为 0~1。它是正负匹配的条件概率的几何平均数的平方。

(19) Phi 4-point correlation

计算公式为

$$PHI(x, y) = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在两项中发生而在另一项中不发生的对角线单元, d 表示事件在两项中都不发生的相应单元。它的取值范围为 -1~1。它是皮尔逊相关系数二值模拟。

(20) Dispersion

计算公式为

$$DISPER(x, y) = \frac{ad - bc}{(a + b + c + d)^2}$$

式中, a 表示事件在两项中都发生的相应单元, b 和 c 表示事件在两项中发生而在另一项中不发生的对角线单元, d 表示事件在两项中都不发生的相应单元。它的取值范围为 -1~1。

10.3.2 距离分析的实例分析

1. 变量间距离分析

例 10.5 仍以例 10.4 中的资料为例, 试对身高、体重、肺活量和胸围变量进行距离分析。

在 SPSS 中对本类型题的具体解题步骤如下:

(1) 在 SPSS 中, 打开 data10-04.sav。

(2) 按 Analyze→Correlate→Distance 顺序单击菜单项, 展开 Distances 距离分析的对话框, 如图 10-6 所示。

在左侧变量名源框中, 选中身高、体重、肺活量和胸围变量, 将它们移入到 Variables 框中。

由于本例的目的是要对身高、体重、肺活量和胸围变量之间进行相似性研究, 在主对话框的 Compute Distances 组中, 选择 Between Variables, 即要求做变量之间的距离分析。因为本例中的 4 个变量都是等间隔变量, 因此, 在 Measure 组中选择 Similarities, 即对变量之间进行相似性测度。

(3) 单击 Measure 按钮，展开如图 10-7 所示的 Similarities Measure 对话框。

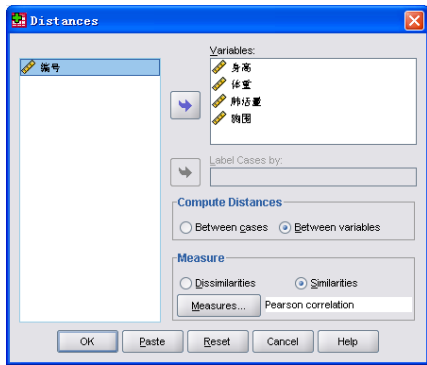


图 10-6 距离分析的对话框

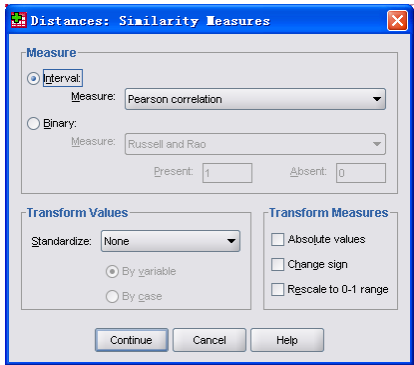


图 10-7 Similarities Measure 对话框

由于本例中全部 4 个变量都是等间隔测度的变量，因此，在 Measure 栏中可以采用系统默认的 Pearson correlation，也可选用 Cosin 选项。由于在例 10.4 中对 4 个变量已做过 Pearson 相关分析，因此，本例选用 Cosin 选项来进行相似性分析。

由于本例中的四个变量在例 10.4 中已得到不拒绝正态分布原假设的统计结论，因此，在 Transform Values 转换数值栏的 Standardized 框中选择 Z scores，对变量做标准化处理，这样做的目的可以消除 4 个变量不同单位量纲对计算结果的影响。同时指定标准化对象为 By variable，即对变量进行标准化。

单击 Continue 按钮，返回距离分析的对话框。

(4) 单击 OK 按钮运行，则在输出窗口中得到计算结果，见表 10-13 和表 10-14。

表 10-13 样品处理汇总

Case Processing Summary					
Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
29	100.0%	0	.0%	29	100.0%

表 10-14 相似系数矩阵

Proximity Matrix				
	Cosine of Vectors of Values			
	身高	体重	肺活量	胸围
身高	1.000	.741	.600	.814
体重	.741	1.000	.751	.834
肺活量	.600	.751	1.000	.805
胸围	.814	.834	.805	1.000

This is a similarity matrix

(5) 结果与分析

表 10-13 中显示的是对样品有效值和缺失值进行的统计。

表 10-14 以矩阵形式给出了变量两两之间的 Cosin 相似系数。相似系数越接近于 1，说明两个变量越相似。

从表 10-14 中可见，体重与胸围变量之间的相似系数最高为 0.834，身高与胸围变量之间的相似系数次之为 0.814，处在第三位的是肺活量与胸围变量之间的相似系数为

0.805, 处在第 4 位的是体重与肺活量之间的相似系数为 0.751, 而体重与身高变量之间的相似系数为 0.741, 处在第 5 位, 肺活量与身高变量之间的相似系数最小, 为 0.805。

2. 样品间距离分析

例 10.6 2002 年世界杯赛, 进入前 16 名的球队在之前的小组赛中的进球与失球数统计见表 10-15, 数据存放在 data10-05.sav 中。试问哪几个队最接近。

表 10-15 进球与失球数统计表

编号	球队名称	进球数	失球数	编号	球队名称	进球数	失球数
1	丹麦	5	2	9	德国	11	1
2	塞内加尔	5	4	10	爱尔兰	5	2
3	西班牙	9	4	11	瑞典	4	3
4	巴拉圭	6	6	12	英格兰	2	1
5	巴西	11	3	13	墨西哥	4	2
6	土耳其	5	3	14	意大利	4	3
7	韩国	4	1	15	日本	5	2
8	美国	5	6	16	比利时	6	5

在 SPSS 中对本类型题的具体解题步骤如下:

(1) 在 SPSS 中, 打开 data10-05.sav。

(2) 按 Analyze→Correlate→Distance 顺序单击菜单项, 展开 Distances 距离分析的对话框, 如图 10-5 所示。

在左侧变量名源框中, 选中进球数和失球数变量, 将它们移入到 Variables 框中。选择球队名称变量将其移入到 Label Cases by 框中, 对样品做出标识。

由于本例的目的是要对 16 支球队(样品)之间进行相似性研究, 因此, 在主对话框的 Compute Distances 组中, 选择 Between Cases, 即要求做样品之间的距离分析。从统计的数据资料中不难看出, 进球数和失球数都是计数资料, 因此, 在 Measure 组中选择 Dissimilarities, 即对变量之间进行相似性测度。

(3) 单击 Measure 按钮, 展开如图 10-6 所示的 Dissimilarities Measure 对话框。

由于本例中的 2 个变量都是计数型资料, 因此, 在 Measure 栏中选择 Counts 选项, 在选择了该选项后, 可以在展开的下拉表中选择两个不相似性测度中的任意一个。本例选用 Chi-square measure 来进行不相似性分析。

其他采用系统默认选项, 不做选择。

单击 Continue 按钮, 返回距离分析的对话框。

(4) 单击 OK 按钮运行, 则在输出窗口中得到计算结果, 见表 10-16 和表 10-17。

表 10-16 样品处理汇总

Case Processing Summary							
Cases							
Valid		Rejected				Total	
		Missing Value		Negative Value			
N	Percent	N	Percent	N	Percent	N	Percent
16	100.0%	0	.0%	0	.0%	16	100.0%

表 10-17 卡方不相似系数矩阵

Proximity Matrix																
	Chi-square between Sets of Frequencies															
	1 丹麦	2 爱尔兰	3 西班牙	4 巴拉圭	5 巴西	6 土耳其	7 韩国	8 美国	9 德国	10 爱尔兰	11 瑞典	12 英格兰	13 墨西哥	14 意大利	15 日本	16 比利时
1 丹麦	.000	.257	.041	.360	.147	.146	.139	.424	.484	.000	.220	.060	.074	.220	.000	.263
2 爱尔兰	.257	.000	.256	.096	.463	.113	.370	.170	.780	.257	.024	.133	.169	.024	.257	.017
3 西班牙	.041	.256	.000	.381	.225	.126	.188	.454	.585	.041	.211	.034	.045	.211	.041	.290
4 巴拉圭	.360	.096	.381	.000	.602	.213	.463	.082	.913	.360	.115	.204	.262	.115	.360	.083
5 巴西	.147	.463	.225	.602	.000	.326	.028	.670	.391	.147	.405	.177	.227	.405	.147	.507
6 土耳其	.146	.113	.126	.213	.326	.000	.270	.282	.653	.146	.062	.051	.064	.082	.146	.135
7 韩国	.139	.370	.188	.463	.028	.270	.000	.520	.286	.139	.334	.170	.201	.334	.139	.395
8 美国	.424	.170	.454	.082	.670	.282	.520	.000	.969	.424	.184	.256	.326	.184	.424	.161
9 德国	.484	.780	.585	.913	.391	.653	.286	.969	.000	.484	.719	.461	.550	.719	.484	.826
10 爱尔兰	.000	.257	.041	.360	.147	.146	.139	.424	.484	.000	.220	.060	.074	.220	.000	.263
11 瑞典	.220	.024	.211	.115	.405	.062	.334	.184	.719	.220	.000	.111	.138	.000	.220	.042
12 英格兰	.060	.133	.034	.204	.177	.051	.170	.256	.461	.060	.111	.000	.000	.111	.060	.148
13 墨西哥	.074	.169	.045	.262	.227	.064	.201	.326	.550	.074	.138	.000	.000	.138	.074	.190
14 意大利	.220	.024	.211	.115	.405	.062	.334	.184	.719	.220	.000	.111	.138	.000	.220	.042
15 日本	.000	.257	.041	.360	.147	.146	.139	.424	.484	.000	.220	.060	.074	.220	.000	.263
16 比利时	.263	.017	.290	.083	.507	.135	.395	.161	.826	.263	.042	.148	.190	.042	.263	.000

This is a dissimilarity matrix

(5) 结果与分析

表 10-16 中显示的是对样品有效值和缺失值进行的统计。

表 10-17 以矩阵形式给出了变量两两之间的卡方不相似系数。不相似系数越近于 0，说明两个变量越接近。

从表 10-17 中可见，丹麦、爱尔兰、日本 3 队之间很接近，卡方不相似系数的计算结果为 0.000，此外，英格兰和墨西哥之间也很接近，卡方不相似系数的计算结果为 0.000。

10.4 典型相关

在 10.1 中，已经对两个变量之间的相关分析作了介绍，但是，当变量有两组而不是两个时，例如，我们在研究肉禽价和粮价之间的关系时，统计了若干年的大豆、玉米、小麦、大麦、红薯及猪肉、牛肉、羊肉、鸡肉、鸭肉等价格，若用上面介绍的两个变量之间的相关分析方法会得出很多个相关系数，我们也许会发现，单独一种粮食的价格和单独一种肉禽价格之间的关系并不密切，所以，很可能从这些相关系数中很难得到两组变量之间的关系，但是长期的生活经验告诉我们，肉禽价和粮价之间非常有关。因此，当研究变量集之间的关系时，不能沿用单变量的研究方法，必须另辟新径。

正是在这种需求下，典型相关分析应运而生。它是用来描述两组随机变量间关系的统计分析方法。它将每组变量综合成一个新变量，这样讨论两个新变量之间的相关关系

时,不但使分析问题的复杂程度简单化,同时还可以达到减少研究变量个数的目的。

将每组变量组合成一个新的综合变量的最简单的方法是对该组变量进行线性组合,但每组变量之间的线性组合有无数多个,因此,必须对其施加一些条件约束,方能使其具有确定性。典型相关分析就是要找到这两组变量线性组合的系数,使得这两个由线性组合生成的变量之间的相关系数最大。因此,典型相关分析的基本思想是,首先从每组变量中找出变量的线性组合,使其具有最大的相关性,然后从每组变量中找出第二对线性组合,使其与第一对线性组合不相关的同时其本身具有最大的相关性,如此下去,直到两组变量间的相关性被提取完毕为止。上述的相关关系用典型相关系数来衡量。解决此类问题的数学工具是矩阵的特征值和特征向量,所得的特征值与典型变量的典型相关系数有直接联系。

10.4.1 典型相关分析的数学模型

设有两组随机变量 $X = (X_1, X_2, \dots, X_p)'$ 和 $Y = (Y_1, Y_2, \dots, Y_q)'$ 的方差分别记作 Σ_{XX} 、 Σ_{YY} , 协方差为 $Cov(X, Y) = \Sigma_{XY}$ 。

令 $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$, 则其协方差阵为 $\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}$

两个系数向量 $a = (a_1, a_2, \dots, a_p)'$ 和 $b = (b_1, b_2, \dots, b_q)'$, 则两组随机变量各自的线性组合,也就是新的综合变量为

$$V = a'X = a_1X_1 + a_2X_2 + \dots + a_pX_p$$

$$W = b'Y = b_1Y_1 + b_2Y_2 + \dots + b_qY_q$$

两个新的典型变量之间的相关系数为

$$Corr(a'X, b'Y) = a'\Sigma_{XX}b / (a'\Sigma_{XX}a b'\Sigma_{YY}b)^{1/2}$$

为了计算确定性,要求在条件

$$D(a'X) = a'\Sigma_{XX}a = 1, \quad D(b'Y) = b'\Sigma_{YY}b = 1$$

下,使 $Corr(a'X, b'Y)$ 最大,其中 $a = a^{(1)}$, $b = b^{(1)}$, 则称 $V_1 = a^{(1)'}X$, $W_1 = b^{(1)'}Y$ 为第一对典型变量,称 $Corr(a'X, b'Y)$ 为第一典型相关系数。

若常数向量 $a = a^{(2)}$, $b = b^{(2)}$, 在条件

$$D(a'X) = a'\Sigma_{XX}a = 1, \quad D(b'Y) = b'\Sigma_{YY}b = 1;$$

$$Cov(V_1, a'X) = 0, \quad Cov(W_1, b'Y) = 0$$

下,使 $Corr(a'X, b'Y)$ 最大,则称 $V_2 = a^{(2)'}X$, $W_2 = b^{(2)'}Y$ 为第二对典型变量,称 $Corr(a'X, b'Y)$ 为第二典型相关系数。

同样,若常数向量 $a = a^{(3)}$, $b = b^{(3)}$, 在条件

$$D(a'X) = a'\Sigma_{XX}a = 1, \quad D(b'Y) = b'\Sigma_{YY}b = 1$$

$$\text{Cov}(V_1, a'X) = 0, \quad \text{Cov}(W_1, b'Y) = 0$$

$$\text{Cov}(V_2, a'X) = 0, \quad \text{Cov}(W_2, b'Y) = 0$$

下, 使 $\text{Corr}(a'X, b'Y)$ 最大, 则称 $V_3 = a^{(3)'}X$, $W_3 = b^{(3)'}Y$ 为第三典型相关变量, 称 $\text{Corr}(a'X, b'Y)$ 为第三典型相关系数。余类推。

根据在条件 $D(a'X) = a'\Sigma_{XX}a = 1$, $D(b'Y) = b'\Sigma_{YY}b = 1$ 下, 使 $\text{Corr}(a'X, b'Y)$ 最大, 可由 Lagrange 乘子法算得

$$Aa = \lambda^2 a, \quad Bb = \lambda^2 b$$

其中, $A = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, $B = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ 。

这说明, λ^2 既是 A 又是 B 的特征根, a 、 b 就是其相应于 A 、 B 的特征向量。

记 A 和 B 的非零特征根和特征向量为

$$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_k^2; \quad a^{(1)}, a^{(2)}, \dots, a^{(k)}; \quad b^{(1)}, b^{(2)}, \dots, b^{(k)}$$

其中 $k = \min(p, q)$, $0 < \lambda_i \leq 1$, 可得 k 对线性组合

$$V_i = a^{(i)'}X, \quad W_i = b^{(i)'}Y, \quad i = 1, 2, \dots, k,$$

每对变量 (V_i, W_i) 称为典型变量。

V_i 与 W_i 之间的相关系数 $\rho_{X_i Y_i} = \text{Corr}(V_i, W_i) = \sqrt{\lambda_i^2} = \lambda_i$ 。因此, 称最大特征值的平方根为第一典型相关系数。

由上述过程可知, 得到的典型变量具有以下两个性质:

1. X 的典型变量 V_1, V_2, \dots 都不相关, Y 的典型变量 W_1, W_2, \dots 也都不相关。
2. λ_i 为 X 和 Y 的同一对典型变量 V_i 和 W_j 之间的相关系数, 不同的 V_i 和 W_j ($i \neq j$) 之间不相关。

在实际问题中, Σ 一般未知, 此时用样本协方差阵或相关阵代替 Σ 就可以了。但此时的特征根可能出现大于 1 的情形, 在这种情况下, 一般统计计算软件中会自动给出调整后的相关系数值。

10.4.2 典型相关系数的检验

在做两组变量 X 和 Y 的典型相关分析之前, 首先应检验两组变量是否相关, 如果不相关, 也就是 $\text{Cov}(X, Y) = 0$, 则讨论两组变量的典型相关就没有任何意义。

设总体 X 的两组随机变量 $X = (X_1, X_2, \dots, X_p)'$ 和 $Y = (Y_1, Y_2, \dots, Y_q)'$, 并且 $X = (X, Y)' \sim N_{p+q}(\mu, \Sigma)$ 。

所要作的原假设 $H_0: \text{Cov}(X, Y) = \Sigma_{12} = 0$

计算

$$\Lambda = \prod_{i=1}^k (1 - \hat{\lambda}_i^2)$$

其中, $\hat{\lambda}_i^2$ 是 $A = R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$ 的特征根, 按大小次序排列为 $\lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_k^2 > 0$, 其中 $k = \min(p, q)$, 当 n 远大于 1 时, 在原假设成立下, 检验统计量为

$$Q_0 = -m \ln \Lambda \sim \chi^2(df)$$

其中, $df = pq$, $m = n - 1 + \frac{1}{2}(p + q + 1)$ 。

因此, 在原假设 $H_0: \text{Cov}(X, Y) = \Sigma_{12} = 0$ 下, 如果计算得到的观测值的显著性水平 $p < 0.05$, 则拒绝原假设, 也就是说第一对典型变量 \hat{V}_1 与 \hat{W}_1 具有相关性, 其相关系数为 $\hat{\lambda}_1$, 即至少可以认为第一个典型相关系数 $\hat{\lambda}_1$ 为显著的。在此前提下, 可以考虑检验其余 $k - 1$ 个典型相关系数的显著性, 此时, 再计算

$$\Lambda_1 = \prod_{i=2}^k (1 - \hat{\lambda}_i^2)$$

则统计量

$$Q_1 = - \left[n - 2 - \frac{1}{2}(p + q + 1) \right] \ln \Lambda_1 \sim \chi_{(p-1)(q-1)}^2$$

因此, 在原假设 $H_0: \text{Cov}(X, Y) = \Sigma_{12} = 0$ 下, 如果计算得到的观测值的显著性水平 $p < 0.05$, 则拒绝原假设, 也就是说第二对典型变量 \hat{V}_2 与 \hat{W}_2 具有相关性, 其相关系数为 $\hat{\lambda}_2$ 。

从以上统计量计算中不难看出有以下的规律, 检验第 r 个 ($r < k$) 典型相关系数的显著性时, 统计量

$$Q_{r-1} = - \left[n - r - \frac{1}{2}(p + q + 1) \right] \ln \Lambda_{r-1} \sim \chi_{(p-r+1)(q-r+1)}^2$$

其中

$$\Lambda_{r-1} = \prod_{i=r}^k (1 - \hat{\lambda}_i^2)$$

因此, 在原假设 $H_0: \text{Cov}(X, Y) = \Sigma_{12} = 0$ 下, 如果计算得到的观测值的显著性水平 $p < 0.05$, 则拒绝原假设, 也就是说第 r 对典型变量 \hat{V}_r 与 \hat{W}_r 具有相关性, 其相关系数为 $\hat{\lambda}_r$ 。

10.4.3 冗余测度

假设每组变量都标准化了, 第一组变量为 $X = (x_1, x_2, \dots, x_p)'$, 第二组变量为 $Y = (y_1, y_2, \dots, y_q)'$, 从第一组变量提取的典型变量为 $V = (v_1, v_2, \dots, v_r)'$, 从第二组变量提

取的典型变量为 $W = (w_1, w_2, \dots, w_r)'$ ， w_i 与 X 分量的相关系数所成向量为 $G_i = (\rho_{i1}, \rho_{i2}, \dots, \rho_{ip})'$ ， v_i 与 Y 分量的相关系数所成向量为 $H_i = (\delta_{i1}, \delta_{i2}, \dots, \delta_{iq})'$ ，则第 i 个典型变量 w_i 从第 1 组变量提取的方差比例为 $G_i'G/p$ ，第 i 个典型变量 v_i 从第 2 组变量提取的方差比例为 $H_i'H/q$ 。

则称 $R_w^{(i)} = \lambda_i G_i'G/p$ ， $R_v^{(i)} = H_i'H/q$ 为冗余测度。

它的大小表示这对典型变量能够对另一组变差相互解释程度的大小，对进一步讨论多对建模提供有用的信息。

10.4.4 实例分析

例 10.7 随机抽取某高中高一男生 38 人，测试反映其体力和运动能力等 12 项指标，测试结果见表 10-18。

体力测试指标： x_1 —反复横向跳（次）、 x_2 —纵跳（cm）、 x_3 —背力（kg）、 x_4 —握力（kg）、 x_5 —台阶试验（指数）、 x_6 —立定体前屈（cm）、 x_7 —俯卧上体后仰（cm）（俯卧背伸测验）。

运动能力测试指标： y_1 —50 米跑（秒）、 y_2 —跳远（cm）、 y_3 —投球（m）、 y_4 —引体向上（次）、 y_5 —耐力跑（秒）。

数据资料已存放在 data10-06.sav 中。试问高中生的体力与运动能力是否相关？

表 10-18 38 名高一男生的 12 项机能、体能和运动能力指标的测试结果

编号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y_1	y_2	y_3	y_4	y_5
1	46	55	126	51	75.0	25	72	6.8	589	27	8	360
2	52	55	95	42	81.2	18	50	7.2	464	30	5	348
3	46	69	107	38	98.0	18	74	6.8	430	32	9	386
4	49	50	105	48	97.6	16	60	6.8	362	26	6	331
5	42	55	90	46	66.5	2	68	7.2	453	23	11	391
6	48	61	106	43	78.0	25	58	7.0	405	27	7	389
7	49	60	100	49	90.6	15	60	7.0	420	21	10	379
8	48	63	122	52	56.1	17	68	7.1	466	28	2	362
9	45	55	105	48	76.0	15	61	6.8	415	24	6	386
10	48	64	120	38	60.2	20	62	7.1	413	28	7	398
11	49	52	100	42	53.4	6	42	7.4	404	23	6	400
12	47	62	100	34	61.2	10	62	7.2	427	25	7	407
13	41	51	101	53	62.4	5	60	8.0	372	25	3	409
14	52	55	125	43	86.3	5	62	6.8	496	30	10	350
15	45	52	94	50	51.4	20	65	7.6	394	24	3	399
16	49	57	110	47	72.3	19	45	7.0	446	30	11	337

(续表)

编号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y_1	y_2	y_3	y_4	y_5
17	53	65	112	47	90.4	15	75	6.6	446	30	12	357
18	47	57	95	47	72.3	9	64	6.6	420	25	4	447
19	48	60	120	47	86.4	12	62	6.8	447	28	11	381
20	49	55	113	41	84.1	15	60	7.0	398	27	4	387
21	48	69	128	42	47.9	20	63	7.1	485	30	7	350
22	42	57	122	46	54.2	15	63	7.2	400	28	6	388
23	54	64	155	51	71.4	19	61	6.9	511	33	12	298
24	53	63	120	42	56.6	8	53	7.5	430	29	4	353
25	42	71	138	44	65.2	17	55	7.0	487	29	9	370
26	46	66	120	45	62.2	22	68	7.4	470	28	7	360
27	45	56	91	29	66.2	18	51	7.9	380	26	5	358
28	50	60	120	42	56.6	8	57	6.8	460	32	5	348
29	42	51	126	50	50.0	13	57	7.7	398	27	2	383
30	48	50	115	41	52.9	6	39	7.4	415	28	6	314
31	42	52	140	48	56.3	15	63	6.9	470	27	11	348
32	48	67	105	39	69.2	23	60	7.6	450	28	10	326
33	49	74	151	49	54.2	20	58	7.0	500	30	12	330
34	47	55	113	40	71.4	19	64	7.6	410	29	7	331
35	49	74	120	53	54.5	22	59	6.9	500	33	21	348
36	44	52	110	34	54.9	14	57	7.5	400	29	2	421
37	52	66	130	47	45.9	14	45	6.8	505	28	11	355
38	48	68	100	45	53.6	23	70	7.2	522	28	9	352

在 SPSS 中的具体解题步骤如下:

1. 找出 canonical correlation.sps (SPSS 中计算典型相关的程序) 存放在你计算机中的具体位置。

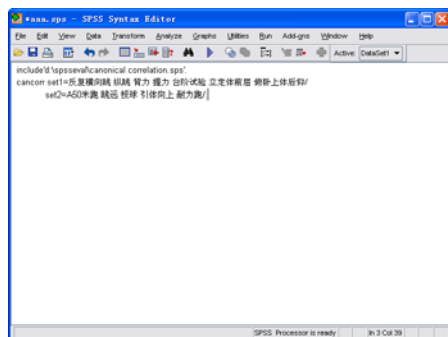


图 10-8 SPSS Syntax Editor

本例中, canonical correlation.sps 存放在计算机 D 盘的 spsseval 文件夹中。

2. 在语句编辑窗口中编写典型相关分析的计算程序语句

在 SPSS 数据编辑窗口中, 先打开 data10-06.sav, 然后按 File→New→Syntax 顺序, (对于已经建立好的语句文件, 则按 File→Open→Syntax 顺序, 直接打开所建立好的文件), 展开 SPSS Syntax Editor, 见图 10-8。

在语句编辑窗口中, 输入下列语句:

include'd:\spsseval\canonical correlation.sps'.
cancorr set1=反复横向跳 纵跳 背力 握力 台阶试验 立定体前屈 俯卧上体后仰/
set2=A50 米跑 跳远 投球 引体向上 耐力跑/.

3. 输出结果及其解释

单击 Run 菜单，选择 ALL（或先选择所要运行的所有语句，再单击工具栏中的 run 图标按钮），则在输出窗口中，得到很长一系列的典型相关的计算结果。为便于分析，我们从上到下将其分成若干张单独的表进行解释，并对输出表格的结构进行了一定的修饰。在输出中由于各变量的输出位数定义为 8 个字节，因此表中的变量与实际变量的对应关系为：

反复横向—反复横向跳，台阶试验—台阶试验指数，立定体前—立定体前屈，俯卧上体—俯卧上体后仰。

表 10-19 给出了第一组变量集中 7 个体力测试指标相互之间的线性相关系数。

表 10-19 第一组变量间的相关系数

Correlations for Set-1							
	反复横向	纵跳	背力	握力	台阶试验	立定体前	俯卧上体
反复横向	1.0000	.3005	.1643	-.0129	.2463	.0722	-.1821
纵跳	.3005	1.0000	.3872	.0253	-.1012	.4561	.2392
背力	.1643	.3872	1.0000	.3153	-.2427	.1931	-.0014
握力	-.0129	.0253	.3153	1.0000	-.0238	.0544	.2092
台阶试验	.2463	-.1012	-.2427	-.0238	1.0000	.0517	.3152
立定体前	.0722	.4561	.1931	.0544	.0517	1.0000	.2801
俯卧上体	-.1821	.2392	-.0014	.2092	.3152	.2801	1.0000

表 10-20 第二组变量间的相关系数

Correlations for Set-2					
	A50 米跑	跳远	投球	引体向上	耐力跑
50 米跑	1.0000	-.4438	-.2574	-.4629	.0745
跳远	-.4438	1.0000	.4389	.5403	-.4249
投球	-.2574	.4389	1.0000	.3598	-.5354
引体向上	-.4629	.5403	.3598	1.0000	-.4200
耐力跑	.0745	-.4249	-.5354	-.4200	1.0000

表 10-20 给出了第二组变量集中 5 个运动能力测试指标相互之间的线性相关系数。

从以上两个表中可见，每组中大多数变量彼此间是弱到中度相关。

表 10-21 给出了第一组变量集中 7 个体力测试指标和第二组变量集中 5 个运动能力测试指标相互之间的线性相关系数。

表 10-21 两组变量中各个变量之间的相关系数

	Correlations Between Set-1 and Set-2				
	A50 米跑	跳远	投球	引体向上	耐力跑
复横向	-.4005	.2939	.4088	.2797	-.4702
纵跳	-.3003	.5253	.5052	.5596	-.2433
背力	-.3026	.5334	.5653	.3215	-.4801
握力	-.2917	.3112	-.0413	.2627	-.1163
台阶试验	-.4295	-.1308	-.0267	.1415	-.0177
立定体前	-.0800	.3196	.2995	.2359	-.2892
俯卧上体	-.2630	.2216	.0402	.0929	.1861

表10-22给出的是5对综合变量之间的典型相关系数，其平方（ λ_i^2 ）可以说明两个变量集之间的共享方差的百分比，由此可知第一典型相关系数说明第一组变量和第二组变量之间的共享方差是 $0.848^2=71.91\%$ 。第一典型相关系数0.848远大于两组变量间的单个相关系数。

表 10-23 给出的是 5 对典型变量之间的典型相关系数等于 0 的显著性检验结果，第一列给出的是 Wilk's Lambda 值，第二列为卡方值，第三列为自由度，第四列为 P 值，从上面的结果中可以看出，第一和第二典型相关系数是显著的（ P 分别为 0.000 和 0.016 小于 0.05），说明第一对和第二对典型变量是有价值的。而第三及以后的典型相关系数不显著（ P 均大于 0.05），所以，第三对及以后的典型变量可以不考虑。

表 10-22 典型相关系数

Canonical Correlations	
1	.848
2	.659
3	.612
4	.400
5	.359

表 10-23 关于剩余相关系数等于 0 的假设检验

Test that remaining correlations are zero:				
	Wilk's	Chi-SQ	DF	Sig.
1	.073	79.877	35.000	.000
2	.259	41.207	24.000	.016
3	.457	23.863	15.000	.067
4	.732	9.525	8.000	.300
5	.871	4.207	3.000	.240

表 10-24、表 10-25 给出的是建立 5 个典型相关的线性组合时，第一组变量中各变量的权重。一般用标准典型系数来建立变量之间的线性组合，它直观上给人以更清楚的印象。由此可得前两个典型变量的线性组合（只有第一和第二典型相关系数是显著的，见表 10-23）

$$v_1 = 0.458x_1 + 0.192x_2 + 0.615x_3 + 0.061x_4 + 0.269x_5 + 0.103x_6 + 0.029x_7$$

$$v_2 = 0.192x_1 - 0.504x_2 + 0.022x_3 + 0.136x_4 + 0.754x_5 - 0.045x_6 - 0.3149x_7$$

式中符号参见题中的说明，上面两式中的各个变量应是标准化处理后的变量。

从中可见，相应于第一个特征值（也是最重要）的典型变量 v_1 主要代表反映躯干背伸肌力量的背力，而相应于第二个特征值（次要）的典型变量 v_2 主要代表反映心血管机

能的台阶试验指数变量。

表 10-24 第一组变量在典型变量中的标准典型系数

	Standardized Canonical Coefficients for Set-1				
	1	2	3	4	5
反复横向	.458	.192	.509	-.398	-.072
纵跳	.192	-.504	-1.050	.580	.094
背力	.615	.022	.479	-.078	.383
握力	.061	.136	-.495	-.252	-.925
台阶试验	.269	.754	-.605	.393	.266
立定体前	.103	-.045	.460	.233	-.454
俯卧上体	.029	-.314	.159	-1.060	.489

表 10-25 第一组变量在典型变量中的一般典型系数

	Raw Canonical Coefficients for Set-1				
	1	2	3	4	5
反复横向	.136	.057	.151	-.118	-.021
纵跳	.028	-.073	-.152	.084	.014
背力	.039	.001	-.031	-.005	.024
握力	.011	.025	-.090	-.046	-.169
台阶试验	.019	.052	-.042	.027	.018
立定体前	.017	-.008	.077	.039	-.076
俯卧上体	.004	-.039	.019	-.130	.060

表10-26、表10-27同表10-24、表10-25，它给出的建立5个典型变量的线性组合时，第二组变量中各变量的权重。由此可得前两个典型变量的线性组合为

$$w_1 = -0.494y_1 + 0.139y_2 + 0.365y_3 + 0.000y_4 - 0.409y_5$$

$$w_2 = -0.898y_1 - 0.875y_2 - 0.483y_3 - 0.267y_4 - 0.756y_5$$

从中可见，（同其他特征值对应的典型变量中的系数相比）相应于第一个特征值（也是最重要）的典型变量 w_1 中各变量系数的绝对值均不大，而且竞赛类项目对应的系数为负，而田赛类项目对应的系数为正，因此，它是反映综合运动能力的变量。相应于第二个特征值（次要）的典型变量 w_2 ，所有变量与其成负面关系，但由于竞赛类项目速度越快，用时越短， w_2 的取值也越大，而田赛类项目成绩越好时， w_2 的取值也越小，因此，它主要代表反映速度运动能力的 50 米跑和耐力跑变量的。

表 10-28 给出的是第一组变量中测试的每个体力指标与第一组变量中五个典型变量之间的典型载荷，典型载荷实际上就是第一组变量中测试的每个体力指标与第一组变量中五个典型变量之间的相关系数。

从表 10-28 可见， v_1 主要和变量背力相关， v_2 主要和变量台阶试验相关。这同在表 10-24 中，它们的典型系数是一致的。

表 10-26 第二组变量在典型变量中的标准典型系数

Standardized Canonical Coefficients for Set-2					
	1	2	3	4	5
A50 米跑	-.494	-.898	.209	.552	-.270
跳远	.139	-.875	.201	-.810	-.499
投球	.365	-.483	.180	.357	1.015
引体向上	.000	-.267	-.976	.791	-.249
耐力跑	-.409	-.756	-.651	-.258	.671

表 10-27 第二组变量在典型变量中的一般典型系数

Raw Canonical Coefficients for Set-2					
	1	2	3	4	5
A50 米跑	-1.409	-2.562	.595	1.575	-.770
跳远	.003	-.018	.004	-.017	-.010
投球	.133	-.176	.065	.130	.370
引体向上	.000	-.070	-.258	.209	-.066
耐力跑	-.013	-.024	-.021	-.008	.022

表 10-28 第一组变量在典型变量中的典型载荷

Canonical Loadings for Set-1					
	1	2	3	4	5
反复横向	.684	.282	.134	.073	-.025
纵跳	.596	-.606	-.414	.236	.080
背力	.739	-.290	.236	-.048	-.037
握力	.259	.042	-.304	-.475	-.730
台阶试验	.226	.742	-.404	-.061	.298
立定体前	.367	-.299	.097	.163	-.242
俯卧上体	.117	-.216	-.351	-.712	.287

表 10-29 给出的是第一组变量中测试的每个体力指标与第二组变量中五个典型变量之间的典交叉载荷，它实际上就是第一组变量中测试的每个体力指标与第二组变量中五个典型变量之间的相关系数。

表 10-29 第一组变量在典型变量中的交叉载荷

Cross Loadings for Set-1					
	1	2	3	4	5
反复横向	.580	.186	.082	.029	-.009
纵跳	.505	-.399	-.254	.094	.029
背力	.626	-.191	.145	-.019	-.013
握力	.220	.028	-.186	-.190	-.262
台阶试验	.191	.489	-.247	-.024	.107
立定体前	.312	-.197	.059	.065	-.087
俯卧上体	.099	-.143	-.215	-.285	.103

表 10-30 给出的是第二组变量中测试的每个运动能力指标与第二组变量中五个典型变量之间的典型载荷，它实际上就是第二组变量中测试的每个运动能力指标与第二组变量中五个典型变量之间的相关系数。

表 10-30 第二组变量在典型变量中的典型载荷

	Canonical Loadings for Set-2				
	1	2	3	4	5
A50 米跑	-.680	-.318	.476	.434	-.144
跳远	.692	-.511	-.063	-.361	-.354
投球	.772	-.327	.212	.282	.416
引体向上	.607	-.180	-.626	.335	-.310
耐力跑	-.700	-.081	-.407	-.396	.425

从表 10-30 可见， w_1 与田赛类变量成正相关，与竞赛类项目成负相关， w_2 与所有变量成负相关。这同在表 10-26 中，它们的典型系数是一致的。

表 10-31 同表 10-29 正相反，表中的数据是第二组五个运动能力变量与第一组五个典型变量之间的相关系数。

表 10-31 第二组变量在典型变量中的交叉载荷

	Cross Loadings for Set-2				
	1	2	3	4	5
A50 米跑	-.576	-.210	.292	.174	-.052
跳远	.587	-.337	-.039	-.144	-.127
投球	.654	-.215	.130	.113	.149
引体向上	.514	-.118	-.383	.134	-.111
耐力跑	-.594	-.053	-.249	-.159	.152

以下输出部分为冗余分析（Redundancy Analysis），它是典型相关的重要内容。

表 10-32 给出了第一组的各典型变量从第一组变量中提取的方差比例，即由第一组变量解释的方差比例。

表 10-32 由第一组变量的各典型变量解释第一组变量的方差的比例

Proportion of Variance of Set-1 Explained by Its Own Can. Var.	
	Prop Var
CV1-1	.234
CV1-2	.174
CV1-3	.090
CV1-4	.118
CV1-5	.110

表 10-33 给出了第一组的各典型变量从第二组变量中提取的方差比例,也即由第二组变量解释的方差的比例。

表 10-33 由第二组变量的典型变量解释第一组变量的方差的比例

Proportion of Variance of Set-1 Explained by Opposite Can. Var.	
	Prop Var
CV2-1	.168
CV2-2	.076
CV2-3	.034
CV2-4	.019
CV2-5	.014

表 10-34 给出了第二组的各典型变量从第二组变量中提取的方差比例,即由第二组变量解释的方差比例。

表 10-34 由第二组变量的典型变量解释第二组变量的方差的比例

Proportion of Variance of Set-2 Explained by Its Own Can. Var.	
	Prop Var
CV2-1	.479
CV2-2	.102
CV2-3	.167
CV2-4	.134
CV2-5	.119

表 10-35 给出了第二组的各典型变量从第一组变量中提取的方差比例,也即由第一组变量解释的方差的比例。

表 10-35 由第一组变量的典型变量解释第二组变量的方差的比例

Proportion of Variance of Set-2 Explained by Opposite Can. Var.	
	Prop Var
CV1-1	.344
CV1-2	.044
CV1-3	.062
CV1-4	.021
CV1-5	.015

结论: 由于 v_1 和 w_1 最相关, 这也说明 v_1 所代表的背力和 w_1 所代表的综合运动能力相关。又因为 v_2 和 w_2 也很相关, 这也说明 v_2 所代表的心血管能力和 w_2 所代表的速度运动能力相关。

注: 在语句编辑窗口中写入以下语句:

manova 反复横向跳 纵跳 背力 握力 台阶试验 立定体前屈 俯卧上体后仰 with

A50 米跑 跳远 投球 引体向上 耐力跑

```
/discrim all alpha(1)
```

```
/print=sig(eigen dim).
```

运行后, 同样也可得到典型相关分析的结果, 还可得到多重回归的结果。

10.5 线性回归分析

10.5.1 线性回归分析概述

10.5.1.1 概念

回归 (regression) 这一术语最早于 1886 年见于 Galton 研究有关遗传显现的文献中。他将“高个子的先代会有高个子的后代, 但后代的增高并不与先代的增高等量”的这一现象称为“向平常高度回归”。虽然, 现在对回归这一概念的理解远非 Galton 的原意, 但这一术语一直沿用至今, 成为统计中最常用的概念之一。

回归分析的目的是寻找一组随机变量 y_i 与另一组随机变量 x_i 的统计依赖关系。这种依赖关系不同于单纯的因果关系, 同函数关系有本质的不同。同相关关系一样, 它在一定的统计意义下是客观存在的。

当作为因变量和自变量的随机变量的个数都为 1 时所作的回归分析称一元回归分析。一元回归分析包括线性和非线性 (曲线回归) 两种。通过对自变量进行一定方式下的数值转换后, 许多时候也可将非线性回归转化成线性回归来处理。

当作为因变量和自变量的随机变量的个数都为多个时所作的回归分析称多对多回归分析。

当作为因变量的个数为 1 时, 而自变量的随机变量的个数为多个时所作的回归分析称多元回归分析。

10.5.1.2 回归分析的主要内容

1. 在同一个总体中随机抽取足量的个体, 对每个个体测试若干项指标, 从而可获得一组数据, 从该数据出发, 确定这些变量间的定量关系式;
2. 对所求得的关系式的可信度进行统计检验;
3. 从影响某一变量的许多变量中, 判断哪些变量的影响是显著的, 哪些变量是不显著的;
4. 利用所求得的关系式对特定的活动或过程进行预报和控制。

此外, 根据回归的分析方法, 特别是进行预报和控制所提出的要求, 选择试验点, 对试验进行某种设计; 寻求点数较少, 且有较好统计性质的回归设计方法, 是近来回归分析研究中的另一个侧重点。

回归分析所研究的数学模型主要是线性回归模型和多项式回归模型。多项式回归模型本身有一些特殊的解决方法，通常也可转化为线性回归模型。

10.5.2 一元线性回归分析

10.5.2.1 数学模型

一元线性回归分析的研究对象是两个呈线性相关关系的变量，通过试验，获取试验数据，从而找出两者之间的经验公式。

表 10-36 10 名男生的百米跑和跳远成绩

编号	100 米跑成绩 (秒)	跳远成绩 (米)
1	12.6	5.43
2	12.3	5.90
3	12.2	5.81
4	11.9	6.00
5	11.8	6.10
6	12.4	5.90
7	12.4	5.65
8	12.5	5.58
9	12.1	5.90
10	11.7	5.99

表 10-36 列出的是通过对 10 名男生的百米跑和跳远成绩进行测试后所获得的数据资料。为研究这些数据中所蕴藏的规律性，我们可以用例 10.1 中介绍的方法，将百米跑成绩作为横坐标，跳远成绩作为纵坐标，在直角坐标系描出各点，见图 10-9，从中可见，这些点大致都在一条直线附近波动，这就是说，变量百米跑和跳远成绩基本上可看作是线性关系，具体要确认这两个变量之间的线性关系，可以用本章 10.1 中介绍的线性相关分析来检验。这些点的偏离是由于试验过程中的其他一些随机因素的

影响所引起的。所以，上表的数据可以假定有以下的结构式

$$y_i = \beta_0 + \beta x_i + \varepsilon_i,$$

$$i = 1, 2, \dots, n$$

其中， $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 分别表示其他随机因素对跳远成绩的影响的总和，称它为随机误差。它需要满足以下几个条件：

条件 1: $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 的数学期望为 0；

条件 2: $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 相互独立，也就是因变量 y 的取值相互独立，它们之间没有联系；

条件 3: $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 有相同的方差 σ^2 ，即方差齐性，也即自变量的任何一个线性组合，因变量 y 的方差应相同；

条件 4: $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 服从同一正态分布 $N(0, \sigma^2)$ ，也就是对于自变量的任何一个线性组合，因变量 y 服从正态分布。

变量 x 可以是随机变量，也可以是一般变量，但通常我们把它看成是可以精确测量或严格控制的变量，即一般变量。

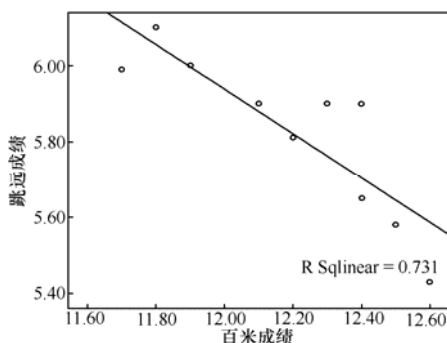


图 10-9 百米跑和跳远成绩的散点图

在满足上述条件下, 变量 y 是服从正态分布 $N(\beta_0 + \beta x_i, \sigma^2)$ 的随机变量。而

$$y_i = \beta_0 + \beta x_i + \varepsilon_i$$

就是一元线性回归的数学模型。其中称 x 为自变量, y 为因变量。 β_0 、 β 为待定参数。

10.5.2.2 参数 β_0, β 的估计

在对 β_0, β 不加任何限制的情况下, 这样的直线可以有无数条, 设 b_0 和 b 分别是参数 β_0 和 β 的一个估计值, 则这无数条直线的一元回归线性方程可写成

$$\hat{y} = b_0 + bx \quad (10.1)$$

式中, b_0 和 b 称为回归方程的**回归系数**, b_0 还叫做回归方程的**截距**, b 又称为回归直线的**斜率**。对于每个 x_i , 由方程 (10.1) 可以确定一个**回归值** \hat{y}_i (又称为**预测值**)。这个预测值 \hat{y}_i 与实测值 y_i 之差 $y_i - \hat{y}_i = y_i - b_0 - bx_i$, 刻画了预测值 \hat{y}_i 与实测值 y_i 的偏离程度。

由不同回归方程组成的直线间, 这种偏离程度是各不相同的。显然, 要判别这些直线间, 到底哪条更适合用来拟合这两个变量间的线性关系, 我们需要引进一个尺度, 很自然的一个想法就是, 对于所有的 x_i , 若预测值 \hat{y}_i 与实测值 y_i 的偏离越小, 就认为直线和所有的

的试验点拟合得越好。但由于偏差的正负值会互相抵消, 会使 $\sum_{i=1}^n (\hat{y}_i - y_i) = 0$, 从而影响到所给出的回归方程的评判, 所以, 我们要用所有预测值 \hat{y}_i 与实测值 y_i 的偏差平方和式 (10.2) 来刻画全部预测值 \hat{y}_i 与实测值 y_i 的偏离程度

$$Q(b_0, b) = \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \sum_{i=1}^N (y_i - b_0 - bx_i)^2 \quad (10.2)$$

显然, 拟合得最好的回归方程应是能使 $Q(b_0, b)$ 是最小的。

在这种要求下, 求解参数 β_0, β 的估计值 b_0 和 b 的方法就是著名的最小二乘法。用它拟合出的直线 $\hat{y} = b_0 + bx$ 和点 (x_i, y_i) , $i = 1, 2, \dots, n$ 的偏离是所有直线中最小的。

根据最小二乘法, 用求极值的原理可以得到下面的方程组

$$\begin{cases} \frac{\partial Q}{\partial b_0} = -2 \sum_{i=1}^N (y_i - b_0 - bx_i) = 0 \\ \frac{\partial Q}{\partial b} = -2 \sum_{i=1}^N (y_i - b_0 - bx_i)x_i = 0 \end{cases} \quad (10.3)$$

解之得

$$\begin{cases} b_0 = \bar{y} - b\bar{x} \\ b = \frac{L_{xy}}{L_{xx}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \end{cases} \quad (10.4)$$

显然, 由此所求得的回归直线是通过点 (\bar{x}, \bar{y}) 的。这一点在作图时用得上。

另一种求回归系数估计的方法是极大似然法, 用该法也同样能得到式 (10.4)。

最小二乘法与极大似然法虽然得到了相同的估计, 但两者是有所不同的, 最小二乘法求回归系数的估计时是不需要正态性假设的, 这一点使得最小二乘法在建立回归模型时, 比极大似然估计应用更加广泛。

理论上可以证明, 当随机误差 ε_i 满足条件 1~3 时, b_0 和 b 分别是 β_0 和 β 的无偏估计, 并且它们的方差也有相同的表达式。

记 b_0 和 b 的方差为 $\sigma_{b_0}^2$ 和 σ_b^2 , 则可以得到

$$\sigma_{b_0}^2 = \left(\frac{1}{n} + \frac{\bar{x}}{L_{xx}} \right) \sigma^2; \quad \sigma_b^2 = \frac{\sigma^2}{L_{xx}}$$

式中, $L_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ (自变量 x 的偏差平方和), σ^2 为误差的方差, n 为样本量, \bar{x} 为自变量 x 的均值。

令 $L_{\text{剩}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, 称为剩余平方和 (也称残差平方和), 则理论上同样可以证明,

$s^2 = \frac{L_{\text{剩}}}{n-2}$ 为误差方差 σ^2 的无偏估计, 也记为 $\hat{\sigma}^2$ 。

利用 s^2 可以得到 $\sigma_{b_0}^2$ 和 σ_b^2 的估计

$$\hat{\sigma}_{b_0}^2 = \left(\frac{1}{n} + \frac{\bar{x}}{L_{xx}} \right) s^2; \quad \hat{\sigma}_b^2 = \frac{s^2}{L_{xx}}$$

它们反映了 b_0 和 b 的估计精度。

从上可知, 回归系数 b 的波动大小不仅与误差的方差有关, 而且还取决于自变量 x 的波动范围的大小, 如果自变量 x 的波动范围较大, 则 b 的波动就较小, 也就是估计比较精确。反之, 如果 x 是在一个较小范围内取得的, 则 β 的估计值就不会精确。回归系数 b_0 的波动不仅与误差的方差和 x 的波动大小有关, 而且还与观测数据的个数 n 有关。数据越多, 且 x 值越分散, 估计 b_0 就越精确。这在安排试验时应引起重视。

10.5.2.3 回归方程的显著性检验

1. 方差分析法

以上, 我们已将回归方程的求解方法作了介绍, 由此可求出回归方程, 但它是否基本上符合变量 y 与 x 之间的客观规律呢? 用它来根据自变量 x 的值来预报因变量 y 的值时效果如何呢? 这就需要对 x 与 y 之间是否是线性关系进行统计检验。

一元线性回归方程的显著性检验, 除可用前面介绍的简单相关系数是否具有显著性意义来检验外, 还可用方差分析的方法来进行检验。

我们知道, 观察值 y_1, y_2, \dots, y_n 之间的差异, 是由两个方面的原因引起的: 一是自变量 x 的取值不同; 另一个是其他因素 (包括试验误差) 的影响。前者可称为条件误差, 后者可称为随机误差。为了检验这两类误差的影响哪个是主要的, 哪个次要的, 首先就必须把它们所引起的差异, 从 y 总的差异中分解出来。

同以前介绍过的方差分析一样, 总偏差平方和记作

$$L_{\text{总}} = L_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (10.5)$$

由于

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \end{aligned}$$

而

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^n (y_i - \hat{y}_i)(b_0 + bx_i - \bar{y}) \\ &= (b_0 - \bar{y}) \sum_{i=1}^n (y_i - \hat{y}_i) + b \sum_{i=1}^n (y_i - \hat{y}_i)x_i \\ &= 0 \end{aligned}$$

(备注: 见方程组 (10.3))

$$\text{因此,} \quad \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (10.6)$$

这就是总偏差平方和的分解公式。

令

$$L_{\text{回}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b^2 L_{xx} = bL_{xy}$$

其中, $L_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = bL_{xx}$, 称 $L_{\text{回}}$ 为回归平方和, 它是由自变量 x 的变化所引起

的, 它的大小 (在与误差相比的意义下) 反映了自变量 x 的重要程度。

令

$$L_{\text{剩}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = L_{yy} - bL_{xy}$$

它是由试验误差以及其他未加控制的因素所引起的, 它的大小反映了试验误差以及其他未加控制的因素对试验结果的影响。

如果变量 y 与变量 x 之间无线性关系, 则模型中的一次项系数 $\beta = 0$; 反之, $\beta \neq 0$ 。所以, 要检验两个变量之间是否有线性关系就是要检验 β 是否为零。理论上可以证明, 在假设 “ $\beta = 0$ ” 成立的条件下, 统计量

$$F = \frac{L_{\text{回}}/1}{L_{\text{剩}}/(n-2)}$$

服从第一自由度为 1 和第二自由度为 $n-2$ 的 F 分布。

当 $F \geq F_{\alpha}(1, n-2)$ 时, 拒绝原假设。说明线性回归模型中的一次项是必要的, 这时我们称该线性方程是显著的。反之, 则称该线性方程是不显著的。它被称为检验回归方程显著性的方差分析。

在线性回归分析中, 还有一个有意义的指标是

$$R^2 = \frac{L_{\text{回}}}{L_{\text{总}}}$$

它表示由于 Y 随 X 的变化而引起的变差与总变差的比值, 也即线性变化引起的变差占 Y 的总变差的比例。它反映用自变量大约可以解释因变量变化的百分比。称其为回归方程的 **决定系数**。 R^2 越接近于 1, 说明 X 对于 Y 的解释作用越强, 回归就越成功。

由于 R^2 有当自变量数目增加而增大的缺点, 因此, 又有一个对其修正的 R^2 , 又称校正 R^2 , 它的意义同 R^2 类似, 其计算公式为

$$\text{Adjusted } R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n-k-1)}{\sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)}$$

式中 k 为自变量的个数, n 为样本容量。

2. 回归系数的假设检验

在对于回归系数 b 检验中, 需要考虑的原假设为

$$H_0: b = b'; H_1: b \neq b'$$

式中, b' 为给定的一个已知值。

在随机误差的 4 个条件满足时, 可证, $b \sim N(\beta, \sigma^2/L_{xx})$, $L_{\text{剩}}/\sigma^2 \sim \chi^2(n-2)$, 于是得到统计量

$$t = \frac{\frac{b-b'}{1/\sqrt{L_{xx}}}}{\sqrt{L_{\text{剩}}/(n-2)}} = \frac{b-b'}{\hat{\sigma}_b}$$

则当原假设 $H_0: b=b'$ 成立时, $t \sim t(n-2)$ 。因此, 当原假设成立的概率 $P(H_0) < \alpha$ (α 为事先给定的检验水平) 时, 拒绝原假设。

同理, 对于 $H_0: b=b'_0; H_1: b_0 \neq b'_0$, 可以得到统计量

$$t = \frac{b_0 - b'_0}{\hat{\sigma}_{b_0}}$$

则当原假设 $H_0: b=b'_0$ 成立时, $t \sim t(n-2)$ 。因此, 当原假设成立的概率 $P(H_0) < \alpha$ (α 为事先给定的检验水平) 时, 拒绝原假设。

同时, 可以得到 b_0 的置信度为 $100(1-\alpha)\%$ 的置信区间为

$$\left[b_0 - \hat{\sigma}_{b_0} t_{\alpha/2}(n-2), b_0 + \hat{\sigma}_{b_0} t_{\alpha/2}(n-2) \right]$$

b 的置信度为 $100(1-\alpha)\%$ 的置信区间为

$$\left[b - \hat{\sigma}_b t_{\alpha/2}(n-2), b + \hat{\sigma}_b t_{\alpha/2}(n-2) \right]$$

在回归模型中, 如果 $b \neq 0$, 称回归方程是显著的, 因此, 在上述检验中, 当所要检验的原假设 $H_0: b=0$ 时, 称它为回归方程的显著性检验。

对于原假设, 我们根据检验结果, 要么拒绝, 要么不拒绝, 无论采取哪种行动, 都可能要犯错误, 拒绝原假设, 可能犯第一类错误, 而不拒绝原假设, 可能要犯第二类错误。

假设原假设为真, 而不拒绝 $H_0: b=0$, 意味着下面的事实之一:

- 线性方程 $\hat{y} = \bar{y} + b(x - \bar{x})$ 对预测 Y 的值与 \bar{y} 的效果一样, Y 的 (条件) 期望与 x 无关, 因而它没有帮助。

- Y 的 (条件) 期望在试验数据的范围内没有线性变化趋势, 但可能是 x 的其他函数形式 (二次函数或 3 次函数等)。

因此, 还应用其他非线性模型作进一步的探索研究。

假设原假设为假, 而拒绝 $H_0: b=0$, 意味着下面的事实之一:

- 用回归方程 $\hat{y} = \bar{y} + b(x - \bar{x})$ 预测 Y 的值时, 比用 \bar{y} 来预测 Y 的值的要好得多, 说明 x 的作用显著。

- 存在 Y 的值随着 x 增加而增加或减少的趋势, 但还不能否定具有这种趋势的 x 的高次项的存在, 因而, 还可能有更好的模型。

例10.8 试对表10-36中列出的10名男生的百米跑成绩和跳远成绩进行一元线性回归分析。数据已存放在data10-07.sav中。

在SPSS中的解题步骤如下：

在数据编辑窗口中，打开data10-07.sav。

1. 首先做出散点图，观察变量间的趋势。

百米成绩和跳远成绩的散点图见图10-9，从该图中可见，这些点大致都在一条直线附近波动，这就是说，百米跑和跳远成绩两个变量基本上可看作是线性关系。

2. 对因变量跳远成绩做正态性检验

用第2章数据资料探索性分析中介绍的方法，可得跳远成绩变量的正态性检验结果，见表10-37。从表中可见，跳远成绩观察的显著性水平为0.115，大于0.05，故现有证据不足以拒绝跳远成绩变量服从正态分布的原假设。

表 10-37 跳远成绩变量的正态性检验

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
跳远成绩	.238	10	.115	.926	10	.409

a. Lilliefors Significance Correction

3. 做一元回归分析

按 Analyze→Regression→Linear 顺序展开 Linear Regression 对话框，见图 10-10。在左侧变量名源框中，选择跳远成绩变量，将其移入到 Dependent 框中，作为因变量，选择百米成绩变量，将其移入到 Independent(s)框中，作为自变量。

其他保持系统默认选择。

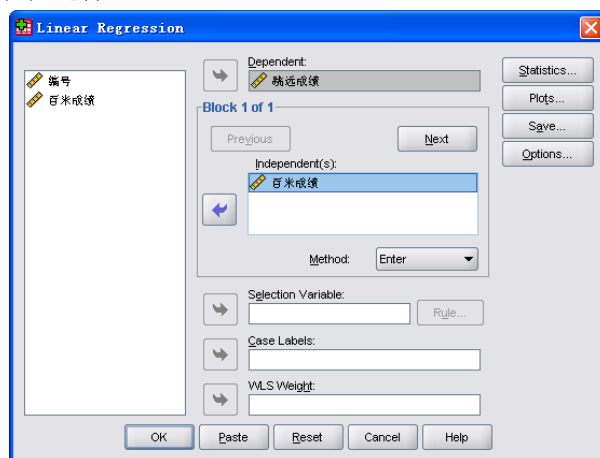


图 10-10 Linear Regression 对话框

单击 OK 按钮运行，在输出窗口中得到 4 张计算结果表，分别参见表 10-38、表 10-39、表 10-40 和表 10-41。

表 10-38 被引入或从方程中剔除的变量

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	百米成绩 ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: 跳远成绩

表 10-39 模型汇总

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.855 ^a	.731	.697	.11565

a. Predictors: (Constant), 百米成绩

表 10-40 方差分析

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.290	1	.290	21.700	.002 ^a
	Residual	.107	8	.013		
	Total	.397	9			

a. Predictors: (Constant), 百米成绩

b. Dependent Variable: 跳远成绩

表 10-41 回归系数及检验

Coefficients ^a						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12.953	1.530		8.464	.000
	百米成绩	-.585	.126	-.855	-4.658	.002

a. Dependent Variable: 跳远成绩

4. 结果与分析

从表 10-38 中可见，只有 1 个模型，被引入方程的变量为百米成绩，它是被强制性引入的（见表中第 4 列方法中采用的 Enter），由于本例是一元线性回归，故没有被剔除的变量。

从表 10-39 中可见，模型 1 的复相关系数 R 为 0.855（在多元线性回归中，称它为复相关系数，由于本例是一元线性回归，所以它也是百米成绩和跳远成绩的简单线性相关系数的绝对值）， R^2 为 0.731，它反映用百米成绩大约可以解释 73.1% 的跳远成绩的变化，校正 R^2 为 0.697，其含义基本同 R^2 的解释，估计的标准误为 0.11585。

从表 10-40 中可见，回归平方和为 0.290，自由度为 1，其方差为 0.290，剩余平方和为 0.107，自由度为 8，其方差为 0.013，因此，计算得到的 F 统计量的值为 21.700， P 值为 0.002，小于 0.01，故拒绝百米成绩和跳远成绩不存在线性相关的原假设，而认为在统计意义上方程是极显著的。

表 10-41 中可见，常数项 b_0 （截距）为 12.953，标准误为 1.530，标准回归系数为 0，检验统计量 $t = 8.464$ ， P 值为 0.000，小于 0.05，说明常数项 b_0 是显著的，同样，回归系数 b （一元回归中的斜率）为 -0.585，标准误为 0.126，标准回归系数为 -0.855，检验统计量 $t = -4.658$ ， P 值为 0.002，小于 0.05，说明回归系数 b 也是显著的，由此，可以说明回归方程是显著的。

在一元线性回归中，表 10-40 和表 10-41 中的检验结果及同 10.1.1 中提到的线性相关系数的检验结果都是等价的。但在多元线性回归中，它们之间并不等价，要分开分析。

根据表 10-41 中得到的回归系数，我们可以写出如下的回归方程

$$\hat{y} = 12.953 - 0.585x$$

10.5.2.4 利用回归方程进行预测和控制

如果回归方程是拟合得好的, 则可用它进一步作预测和控制。

1. 预测

对任一给定的 x_0 , 利用回归方程可以得到 $\hat{y}_0 = b_0 + bx_0$, 它是随机变量 y_0 的一个预测值, 称为点预测。

在实际应用中, 常常要求在 $x = x_0$ 时, 对相应的随机变量 y_0 的取值范围作预测, 称这种预测为区间预测, 又称预报。区间预测类似于一个区间估计问题, 即在一定的显著性水平 α 下, 寻找一个正数 δ , 使得实际观察值 y_0 以 $1-\alpha$ 的概率落在区间 $(\hat{y}_0 - \delta, \hat{y}_0 + \delta)$ 内, 也就是 $P\{|y_0 - \hat{y}_0| \leq \delta\} = 1 - \alpha$ 。

由于

$$\frac{y_0 - \hat{y}_0}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}}} \sim t(n-2)$$

所以, 理论上可以证明,

$$\delta = t_{1-\alpha/2}(n-2) \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}} \right]}$$

上式表明, 利用回归方程预测实际观察值 y_0 的偏差 δ 与显著性水平 δ 有关, δ 越小, $t_{1-\alpha/2}(n-2)$ 越大, δ 就越大, 同样与 \sqrt{n} 的大小也有关系, n 越大 δ 就越小, 同时还与 x_0 有关, x_0 离 \bar{x} 越近, δ 越小, 离 \bar{x} 越远, δ 越大。

由上可得, y_0 的置信度为 $100(1-\alpha)\%$ 的预测区间为

$$\left[\hat{y}_0 - t_{1-\alpha/2}(n-2) \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}} \right]}, \hat{y}_0 + t_{1-\alpha/2}(n-2) \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}} \right]} \right]$$

当把上面提到的 y_0 的置信度为 $100(1-\alpha)\%$ 的预测区间公式中的 x_0 代之以任意给定的 x ,

记 $\sigma(x) = t_{1-\alpha/2}(n-2) \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{L_{xx}}}$, 则对应 y 夹在以下两条曲线

$$\hat{y}_1 = \hat{y} - \delta(x) = b_0 + bx - \sigma(x)$$

$$\hat{y}_2 = \hat{y} + \delta(x) = b_0 + bx + \sigma(x)$$

之间。

假如分别做出函数 $\hat{y}_1 = \hat{y} - \delta(x)$ 和 $\hat{y}_2 = \hat{y} + \delta(x)$ 的图形, 则它们分别对称地在直线 $\hat{y} = b_0 + b_1x$ 的上方和下方两侧, 两头呈喇叭形。从图形上而言, 过 x 的纵坐标线夹于 y_1 和 y_2 两曲线间的一段就是 y 的预测区间。

在不重复试验的情况下, $\frac{L_{\text{剩}}}{n-2}$ 是 σ^2 的无偏估计, 即 $\sigma^2 = \frac{L_{\text{剩}}}{n-2}$

在重复试验的情况下, $\frac{L_{\text{误}}}{n(m-1)}$ 是 σ^2 的无偏估计, 即 $\sigma^2 = \frac{L_{\text{误}}}{n(m-1)}$

在重复试验场合, 当用误差平方和检验失拟平方和的结果不显著时, $\frac{L + L_{\text{失}}}{nm-2}$ 是 σ^2 的无偏估计, 即 $\sigma^2 = \frac{L_{\text{误}} + L_{\text{失}}}{nm-2}$ 。这样, 我们可以作预测了。

2. 控制

所谓控制问题实质上是预报的反问题, 它所涉及的是当观察值 y 在 $y_1 < y < y_2$ 范围内取值时, 应把自变量 x 控制在什么样的范围内取值。即要寻找两个数 x_1, x_2 , 使得

$$\hat{y} - \delta(x_1) > y_1$$

$$\hat{y} + \delta(x_2) < y_2$$

当 n 比较大且 x_0 在 \bar{x} 附近取值时, 有

$$1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2} \approx 1$$

又因为

$$\sigma^2 \approx \hat{\sigma}^2 = \frac{L_{\text{剩}}}{n-2}$$

所以利用正态分布性质有

$$P\{\hat{y}_0 - 2\hat{\sigma}\langle y_0 \rangle \hat{y}_0 + 2\hat{\sigma}\} = 95\%$$

$$P\{\hat{y}_0 - 3\hat{\sigma}\langle y_0 \rangle \hat{y}_0 + 3\hat{\sigma}\} = 99\%$$

因此, 如要控制 y 有 95% 的把握在 $y_1 < y < y_2$ 内取值, 只要通过方程

$$y_1 = b_0 - 2\hat{\sigma} + bx_1$$

$$y_2 = b_0 + 2\hat{\sigma} + bx_2$$

分别解出 x_1 和 x_2 , 从而确定 x 值的控制范围。

例 10.9 对某产品进行腐蚀刻线试验, 得腐蚀深度 y 与腐蚀时间 x 的数据见表 10-42。数据已存放在 data10-08.sav 中。

表 10-42 腐蚀深度 y 与腐蚀时间 x 的数据

x (s)	5	10	15	20	30	40	50	60	70	90	120
y (μm)	6	10	10	13	16	17	19	23	25	29	46

(1) 求 y 关于 x 的线性回归方程, 并对线性回归方程进行检验;

- (2) 求 $x_0 = 75$ 时 \hat{y}_0 的 95% 的预测区间;
- (3) 若要求腐蚀深度在 $10 \sim 20 \mu\text{m}$ 之间, 试问 x 应控制在何范围内?
- 在 SPSS 中的解题步骤如下:

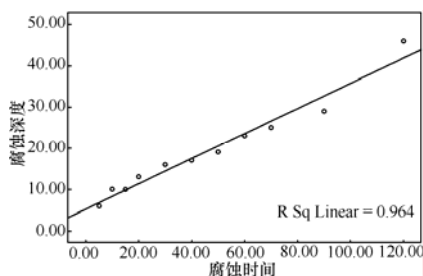


图 10-11 腐蚀深度与腐蚀时间的散点图

在数据编辑窗口中, 打开 data10-08.sav。

1. 首先做出散点图, 观察变量间的趋势。

用例 10.1 中介绍的方法, 将腐蚀时间作为横坐标, 腐蚀深度作为纵坐标, 在直角坐标系描出各点, 见图 10-11, 从中可见, 这些点大致都在一条直线附近波动, 这就是说, 变量腐蚀时间和腐蚀深度基本上可看作是线性关系。

2. 对因变量腐蚀深度做正态性检验

用第 2 章数据资料探索性分析中介绍的方法, 可得腐蚀深度变量的正态性检验结果, 见表 10-43。从表中可见, 腐蚀深度观察的显著性水平为 0.200 以上, 大于 0.05, 故现有证据不支持拒绝腐蚀深度变量服从正态分布的原假设。

表 10-43 腐蚀深度变量的正态性检验

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
腐蚀深度	.153	11	.200 [*]	.904	11	.206

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

3. 做一元回归分析

按 Analyze→Regression→Linear 顺序展开 Linear Regression 对话框, 见图 10-10。在左侧变量名源框中, 选择腐蚀深度变量, 将其移入到 Dependent 框中, 作为因变量, 选择腐蚀时间变量, 将其移入到 Independent(s)框中, 作为自变量。

其他保持系统默认选择。

单击 OK 按钮运行, 同例 10.8 一样, 可在输出窗口中得到 4 张计算结果表, 分别参见表 10-44、表 10-45、表 10-46 和表 10-47。

表 10-44 被引入或从方程中剔除的变量

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	腐蚀时间 ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: 腐蚀深度

表 10-45 模型汇总

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.982 ^a	.964	.960	2.23559

a. Predictors: (Constant), 腐蚀时间

表 10-46 方差分析

ANOVA ^a					
Model	Sum of Squares	df	Mean Square	F	Sig.
1					
Regression	1213.746	1	1213.746	242.852	.000 ^a
Residual	44.981	9	4.998		
Total	1258.727	10			

a. Predictors: (Constant), 腐蚀时间
b. Dependent Variable: 腐蚀深度

表 10-47 回归系数及检验

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1					
(Constant)	5.344	1.129		4.735	.001
腐蚀时间	.304	.020	.982	15.584	.000

a. Dependent Variable: 腐蚀深度

对这 4 个表的解释同对表 10-38~表 10-41 的解释。

从表 10-47 中可得，所求的腐蚀时间对腐蚀深度的一元线性回归方程为

$$\hat{y} = 5.344433 + 0.3043357x$$

注：在输出窗口中，双击上述两个回归系数的值，可加大该单元格宽度，以显示更多位小数。

4. 方程的显著性检验

从表 10-46 中可知，在“ $\beta = 0$ ”的原假设下，出现目前统计量的值或更极端值的概率为 0.000，小于 0.05，故拒绝原假设，即认为上述的一元线性回归方程在统计上是有显著性意义的。表 10-47 中对回归系数的显著性检验，得到同方差分析一样的统计结论。

5. 求 $x_0 = 75$ 时 \hat{y}_0 的 95% 的预测区间

将 $x_0 = 75$ 代入上面求到的预测方程中，得到

$$\hat{y}_0 = 5.344433 + 0.3043357 \times 75 = 28.17$$

由于回归方程标准差

$$\hat{\sigma} = s = \sqrt{\frac{L_{\text{剩}}}{n-2}} = 2.23559$$
$$L_{\text{回}} = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2 = b^2 L_{xx} = 1213.74635$$
$$b = 0.3043357$$

所以

$$L_{xx} = \frac{1213.746}{0.3043357^2} = 13104.55$$

而

$$t_{1-0.05/2}(9) = IDF.T(0.975, 9) = 2.2622$$

式中 $IDF.T(P, df)$ 为 T 分布的反函数，用它可求已知 T 分布左侧概率 P 和自由度 df 下对应的 t 值。

用描述统计的方法，可得

$$\bar{x} = 46.3636$$

故

$$\delta = 2.6222 \times 2.23559 \times \sqrt{1 + \frac{1}{11} + \frac{(75 - 46.3636)^2}{13104.55}} = 5.431634$$

因此, y_0 的置信度为 $100(1-\alpha)\%$ 的预测区间为

$[28.17 - 5.431634, 28.17 + 5.431634]$, 即 $[22.74, 33.60]$ 。

6. 求腐蚀深度在 $10 \sim 20\mu\text{m}$ 之间时 x 控制的范围

当 $y_1 = 10$, $y_2 = 20$ 时, 也即 y 的取值区间为 $[10, 20]$ 时, 由方程组

$$\begin{cases} y_1 = b_0 + bx_1 - 2\hat{\sigma} \\ y_2 = b_0 + bx_2 + 2\hat{\sigma} \end{cases} \text{ 得 } \begin{cases} x_1 = \frac{y_1 + 2\hat{\sigma} - b_0}{b} \\ x_2 = \frac{y_2 - 2\hat{\sigma} - b_0}{b} \end{cases}$$

$$\text{将上述已知条件代入得 } \begin{cases} x_1 = \frac{y_1 + 2\hat{\sigma} - b_0}{b} = \frac{10 + 2 \times 2.23559 - 5.344433}{0.3043357} = 29.99 \\ x_2 = \frac{y_2 - 2\hat{\sigma} - b_0}{b} = \frac{20 - 2 \times 2.23559 - 5.344433}{0.3043357} = 33.46 \end{cases}$$

即当腐蚀时间控制在 29.99 秒到 33.46 秒之间时, 腐蚀深度大约在 $10 \sim 20\mu\text{m}$ 之间。

10.5.2.5 模型检查

上面提到的估计和假设检验是以一元线性回归模型及其 4 个基本条件的满足为基础的。所以, 这些假定如果出现问题, 则相应的推断也就会受到影响。因此, 利用观测数据检验模型的假定是否存在明显问题是很有必要的。

对模型误差和方差齐性, 即条件 1 和条件 3 的假定进行检验时, 需要观测数据资料中有重复测量数据, 因此, 进行全部或部分试验点的重复试验是检验这两个假定的关键。

1. 重复试验下模型误差的检验

上面所做出的“回归方程显著”这一判断是建立在用剩余平方和去检验回归平方和的基础上的, 它只是表明对于其他因素和试验误差来说, 因素 x 的一次项对指标 y 的影响是主要的, 但它并没有告诉我们: 除 x 以外, 是否还有一个或几个不可忽略的其他因素影响 y , 以及 x 和 y 的关系是否确是线性关系。换言之, 上述的“回归方程显著”, 并不表明这个回归方程是拟合得很好的。

要检验一个回归方程拟合得好坏, 还需要做一些重复试验。通过重复试验可以将剩余平方和作进一步分解, 从而获得误差平方和与失拟平方和, 用误差平方和对失拟平方和进行 F 检验, 就可确定回归方程拟合程度的好坏。

重复试验可以对全部试验点进行, 也可对部分试验点进行。

(1) 对部分试验点进行重复试验时, 又可对一个或几个试验点进行重复。为讨论方便, 我们假设仅对第 n 号试验进行了 m 次重复, 得到 $n+m-1$ 个数据

$$y_1, y_2, \dots, y_{n-1}, y_n, y_{n+1}, \dots, y_{n+m-1}$$

其中前 $n-1$ 个试验没有重复, 后 m 个试验是重复第 n 号试验。记这 $n+m-1$ 个数据的均值为 \bar{y} , 对这 $n+m-1$ 个数据可以用上面一样的方法来算得回归系数和各种平方和

$$\begin{aligned} L_{\text{总}} = L_{yy} &= \sum_{i=1}^{n+m-1} (y_i - \bar{y})^2, f_{\text{总}} = (n+m-1) - 1 \\ L_{\text{回}} &= \sum_{i=1}^{n+m-1} (\hat{y}_i - \bar{y})^2 = b^2 L_{xx} = b L_{xy}, f_{\text{回}} = 1 \\ L_{\text{剩}} &= \sum_{i=1}^{n+m-1} (y_i - \hat{y})^2 = L_{yy} - b L_{xy}, f_{\text{剩}} = N + m - 3 \end{aligned}$$

而误差平方和可用后面的 m 个数据算得

$$L_{\text{误}} = \sum_{i=n}^{n+m-1} (y_i - \bar{y}_N)^2, f_{\text{误}} = m - 1$$

其中, \bar{y}_N 是 $y_N, y_{N+1}, \dots, y_{N+m-1}$ 的均值。

由于剩余平方和反映了试验误差以及其他未控因素对试验结果的影响, 故, 假如从 $L_{\text{剩}}$ 中除去 $L_{\text{误}}$, 则剩余下来的 $L_{L_f} = L_{\text{剩}} - L_{\text{误}}, f_{L_f} = (n+m-3) - (m-1) = n-2$, 就只反映了其他未控因素的影响, 即回归拟合得好坏的程度, 所以, L_{L_f} 称为失拟平方和。这时总的偏差平方和可用下式表示

$$L_{\text{总}} = L_{\text{回}} + L_{L_f} + L_{\text{误}} \quad (10.7)$$

可以证明, 在原假设 $H_0: \beta = 0$ 为真时, 统计量

$$F_1 = \frac{L_{L_f} / f_{L_f}}{L_{\text{误}} / f_{\text{误}}}$$

服从自由度为 f_{L_f} 和 $f_{\text{误}}$ 的 F 分布。用它可检验回归方程拟合得是好还是坏。

对于给定的显著性水平 α , 当 $p(H_0) \geq \alpha$, 则第一次 F 检验结果不显著, 这说明失拟平方和基本上是由试验误差等偶然因素引起的。此时可把 L_{L_f} 和 $L_{\text{误}}$ 合并, 并用合并值来检验 $L_{\text{回}}$, 即

$$F_2 = \frac{L_{\text{回}} / f_{\text{回}}}{(L_{L_f} + L_{\text{误}}) / (f_{L_f} + f_{\text{误}})}$$

它服从第一自由度为 $f_{\text{回}}$, 第二自由度为 $f_{L_f} + f_{\text{误}}$ 的 F 分布。

当第二次 F 检验的结果显著, 那么就称回归方程是拟合得好的; 如果第二次 F 检验的结果不显著, 此时有两种可能: 一种可能是没有什么因素对 y 有系统影响; 另一种可能是试验误差过大。显然, 此时所求得的回归方程是不理想的。

当第一次 F 检验结果显著 ($p(H_0) < \alpha$) 时, 说明影响失拟平方和的除试验误差外,

还有其他一些因素。这时有几种可能：第一种可能是除 x 影响 y 外，至少还有一个不可忽略的因素影响 y ；第二种可能是 y 和 x 为曲线关系；此外，另一种可能是 y 和 x 无关。

在第一次 F 检验结果显著时，即使用 $L_{\text{误}}$ 对 $L_{\text{回}}$ 进行第二次 F 检验的结果也显著，此时，所得到的一元线性回归方程，只能说有一定的作用，但已不能说它是拟合得好的，要给出好的结论，还需要查明原因，改变数学模型 (10.1) 作进一步研究后方能得出。

(2) 在对全部试验点进行重复试验时的参数估计和统计检验。

假设对全部 n 个试验点各进行 m 次重复试验，总共获得 $n \cdot m$ 个数据，则在重复试验情况下一元线性回归的数学模型为

$$\begin{aligned} y_{ij} &= \beta_0 + \beta x_i + \varepsilon_{ij}, \\ i &= 1, 2, \dots, n, j = 1, 2, \dots, m \end{aligned}$$

其中， ε_{ij} 是一组相互独立的服从 $N(0, \sigma^2)$ 的随机变量。

在重复试验情况下，用最小二乘法同样可算得参数 β_0, β 的最小二乘估计为

$$\begin{cases} b_0 = \bar{y} - b\bar{x} \\ b = \frac{L_{x\bar{y}_i}}{L_{xx}} = \frac{\sum x\bar{y}_i - \frac{\sum x \sum \bar{y}_i}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \end{cases} \quad (10.8)$$

$$\text{其中, } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \bar{y} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m y_{ij}, \bar{y}_i = \frac{1}{m} \sum_{j=1}^m y_{ij}$$

式 (10.8) 同式 (10.4) 相比可得出结论，用每个试验点上的平均观察值拟合出的回归方程与用原来 $n \cdot m$ 个观察值拟合出来的回归方程是完全一样的。

但需要提醒的是，在求以下讨论的总的偏差平方和时，用包括重复测试的全部原始数据和用每个试验点上的平均观察值所求的总的偏差平方和是不一样的。因此，它会影响到数据文件的建立的格式。数据文件建立的格式可参见 data10-09.sav。

同上一样，总的偏差平方和可分解为

$$L_{\text{总}} = L_{\text{回}} + L_{L_f} + L_{\text{误}}$$

其中

$$\begin{aligned} L_{\text{总}} &= L_{yy} = \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \bar{y})^2 = \sum_{i=1}^n \sum_{j=1}^m y_{ij}^2 - \frac{\left(\sum_{i=1}^n \sum_{j=1}^m y_{ij} \right)^2}{nm}, f_{\text{总}} = nm - 1 \\ L_{\text{回}} &= \sum_{i=1}^n \sum_{j=1}^m (\hat{y}_i - \bar{y})^2 = mb^2 L_{xx} = mb L_{x\bar{y}_i}, f_{\text{回}} = 1 \end{aligned}$$

$$L_{\text{误}} = \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2, f_{\text{误}} = n(m-1)$$

$$L_{L_f} = \sum_{i=1}^n \sum_{j=1}^m (\bar{y}_i - \hat{y}_i)^2 = m \sum_{i=1}^n (\bar{y}_i - \hat{y}_i)^2, f_{L_f} = n-2$$

可以证明, 在原假设 $H_0: \beta=0$ 为真时, 统计量

$$F_1 = \frac{L_{L_f} / f_{L_f}}{L_{\text{误}} / f_{\text{误}}}$$

服从自由度为 f_{L_f} 和 $f_{\text{误}}$ 的 F 分布。

同上面的判定方法一样, 当第一次 F 检验结果显著 ($p(H_0) < \alpha$) 时, 说明影响失拟平方和的除试验误差外, 还有其他一些因素。需进一步查明原因, 再作研究; 而当第一次 F 检验结果不显著时, 可把 f_{L_f} 和 $L_{\text{误}}$ 合并, 并用合并值来检验 $L_{\text{回}}$,

$$F_2 = \frac{L_{\text{回}} / f_{\text{回}}}{(L_{L_f} + L_{\text{误}}) / (f_{L_f} + f_{\text{误}})} \sim F(1, nm-2)$$

当 $p(H_0) < \alpha$ 时, 说明回归方程是显著的, 在这种条件下得出的“回归方程显著”才表示回归方程拟合得是较好的。而当 $p(H_0) > \alpha$, 即第二次 F 检验的结果不显著时, 则说明把 x 的一次项引入回归方程没有多大作用, 这时可能是由于试验误差过大, 或对 y 有影响的因素可能不存在。

例 10.10 从某产品中随机抽取两件, 对其进行腐蚀刻线重复试验, 得腐蚀深度 y 与腐蚀时间 x 的数据见表 10-48。数据已存放在 data10-09.sav 中。

表 10-48 重复测定的腐蚀深度 y 与腐蚀时间 x 的数据

x (s)	5	10	15	20	30	40	50	60	70	90	120
y_1 (μm)	6	10	10	13	16	17	19	23	25	29	46
y_2 (μm)	7	10	11	13	15	16	19	24	25	30	45

求 y 关于 x 的线性回归方程, 并进行显著性检验。

在 SPSS 中的解题步骤如下:

在数据编辑窗口中, 打开 data10-09.sav。

① 首先做出散点图, 观察变量间的趋势。

用例 10.1 中介绍的方法, 将腐蚀时间作为横坐标, 腐蚀深度作为纵坐标, 在直角坐标系描出各点, 见图 10-12, 从中可见, 这些点大致都在一条直线附近波动, 这就是说, 变量腐蚀时间和腐

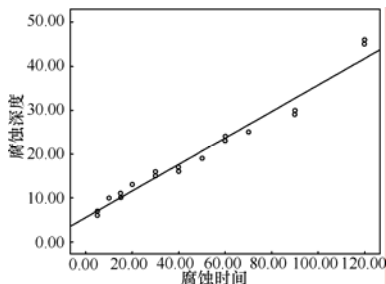


图 10-12 腐蚀深度与腐蚀时间的散点图

蚀深度基本上可看作是线性关系。

②对因变量腐蚀深度做正态性检验

用第 2 章数据资料探索性分析中介绍的方法,可得腐蚀深度变量的正态性检验结果,见表 10-49。从表中可见,腐蚀深度观察的显著性水平为 0.184,大于 0.05,故现有证据不足于拒绝腐蚀深度变量服从正态分布的原假设。

表 10-49 腐蚀深度变量的正态性检验

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
腐蚀深度	.155	22	.184	.887	22	.016

a. Lilliefors Significance Correction

③做一元回归分析

按 Analyze→Regression→Linear 顺序展开 Linear Regression 对话框,见图 10-10。在左侧变量名源框中,选择腐蚀深度变量,将其移入到 Dependent 框中,作为因变量,选择腐蚀时间变量,将其移入到 Independent(s)框中,作为自变量。

其他保持系统默认选择。

单击 OK 按钮运行,同例 10.8 一样,可在输出窗口中得到 4 张计算结果表,分别参见表 10-50、表 10-51、表 10-52 和表 10-53。

表 10-50 被引入或从方程中剔除的变量

Variables Entered/Removed ^b			
Model	Variables Entered	Variables Removed	Method
1	腐蚀时间 ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: 腐蚀深度

表 10-51 模型汇总

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.984 ^a	.969	.968	1.95234

a. Predictors: (Constant), 腐蚀时间

表 10-52 方差分析

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2387.267	1	2387.267	626.309	.000 ^a
	Residual	76.233	20	3.812		
	Total	2463.500	21			

a. Predictors: (Constant), 腐蚀时间

b. Dependent Variable: 腐蚀深度

表 10-53 回归系数及检验

Coefficients ^a						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.507	.697		7.901	.000
	腐蚀时间	.302	.012	.984	25.026	.000

a. Dependent Variable: 腐蚀深度

对这 4 个表的解释同对表 10-38~表 10-41 的解释。

从表 10-53 中可得,所求的腐蚀时间对腐蚀深度的一元线性回归方程为

$$\hat{y} = 5.507 + 0.302x$$

④方程的显著性检验

由于本例是重复测试情形，因此，首先要进行 F_1 检验，故需将剩余平方和分解为误差平方和与失拟平方和。这可以直接用公式计算而得，也可以利用 SPSS 中现成的一般线性模型中的单因变量方差分析程序来获得误差平方和。

具体做法参见第 8 章多因素方差分析。由此可得表 10-54。

表 10-54 单因变量方差分析

Tests of Between-Subjects Effects					
Dependent Variable: 腐蚀深度					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2460.000 ^a	10	246.000	773.143	.000
Intercept	8365.500	1	8365.500	2.629E4	.000
腐蚀时间	2460.000	10	246.000	773.143	.000
Error	3.500	11	.318		
Total	10829.000	22			
Corrected Total	2463.500	21			

a. R Squared = .999 (Adjusted R Squared = .997)

从表中可见， $L_{\text{误}} = 3.5$, $f_{\text{误}} = n(m-1) = 11 \times (2-1) = 11$ ，误差的方差为 0.318。

从表 10-52 可得， $L_{\text{回}} = 2387.267$ ， $f_{\text{回}} = 1$ ，回归方差为 2387.267。由于 $L_{\text{剩余}} = 76.233$ ，所以， $L_{L_f} = L_{\text{剩余}} - L_{\text{误}} = 76.233 - 3.5 = 72.733$ ， $f_{L_f} = n - 2 = 11 - 2 = 9$

$$\text{由此可得， } F_1 = \frac{L_{L_f} / f_{L_f}}{L_{\text{误}} / f_{\text{误}}} = \frac{72.733/9}{3.5/11} = 25.39883$$

利用第 2 章中介绍的计算派生指标的方法，结合 $\text{CDF.F}(\text{quant}, df1, df2)$ 函数计算给定第一自由度 $df1$ 和第二自由度 $df2$ 下，小于数值 quant 的累积概率值的特性，用 $1-\text{CDF.F}(25.39883, 9, 11)$ ，可得到大于等于 25.39883 的 F 分布的右侧概率为 0.0000，由于它小于 0.05，说明失拟平方和中除含有试验误差的影响外，尚有其他因素的影响，需进一步查明原因，再作研究。

当 F_1 检验结果不显著时，才需要进行 F_2 检验，而 F_2 检验就是表 10-52 中的方差分析结果。

重复试验对统计分析是有好处的，但重复试验会受到很多条件的限制，如时间、设备、经费等，有时它几乎是不可能的，如测量相同条件下的气象数据。在这种条件下，回归方程是否拟合得好坏的信息就不能用统计检验的方法来获得，只有靠往后的实践直接进行检验了。

2. 重复试验下方差齐性的检验

随机误差需要满足的条件3方差齐性假定各次观测是同方差的，如果误差仅仅是由测量误差组成的，则这一假定相当于所有测量的精度是一致的，也相当于 y 的方差是常量。当在 x_i 处重复观测的次数 $n_i > 1$ 时，每个点处都有反映该点处随机误差方差的信息，此时

就能进行方差齐性假定的检验。

记 $\text{Var}(y_{ij}) = \sigma_i^2, (i=1, 2, \dots, m)$, 其中 m 为自变量 x 所观测的点数, 在条件 2 随机误差相互独立成立时, 要检验的假设 $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_m^2$; $H_1: \sigma_i^2$ 不全相等。

Levine 的检验方法为:

记 \tilde{y}_i 为样本 $y_{i1}, y_{i2}, \dots, y_{in_i}$ 的中位数

$$z_{ij} = |y_{ij} - \tilde{y}_i|; \quad \bar{z}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij}; \quad \bar{z}_{..} = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} z_{ij}$$

其中 n 为样本量, 则 Levine 检验的统计量为

$$L = \frac{\sum_{i=1}^m n_i (\bar{z}_{i.} - \bar{z}_{..})^2 / (m-1)}{\sum_{i=1}^m \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_{i.})^2 / (n-m)}$$

对于给定显著性水平 α , 当 $L \geq F_{1-\alpha}(m-1, n-m)$ 时拒绝 H_0 , 认为各个方差不都相同。本检验不需要正态性假设。

例10.11 某种合金钢的抗拉强度 y_1 (单位: 千克/毫米²) 和延伸率 y_2 (单位: %) 与钢中含碳量 x 有一定关系。实测 68 炉钢样的含碳量 x 、抗拉强度 y_1 和延伸率 y_2 的数据, 已存放在数据文件 data10-10.sav 中, 试分别对抗拉强度 y_1 及延伸率 y_2 在含碳量 x 各点处的方差进行齐性检验。

【题析】 数据文件中, 钢中的含碳量数值相同时, 如都是 0.05 (共出现 4 次), 相当于在含碳量为 0.05 时, 重复了四次试验, 同理当含碳量为 0.08 时, 共出现 7 次, 相当于在含碳量为 0.08 时, 重复了七次试验, \dots , 其余类推, 共有 10 个不同的含碳量值, 因此, 本例可看作在自变量 x_i 处重复观测的次数为 n_i 时, 每个点处都有反映该点处随机误差方差的信息, 所以, 此时可以进行方差齐性假定的检验。

需要注意的是, 当各点重复测试次数的最小值小于 4 时, 用 Levine 中位数检验法进行方差齐性检验时, 需要手工方法进行, 而当各点重复测试次数的最小值大于等于 4 时, 可用第 2 章 2.4 中提到过的探索性分析中的 Levine 中位数检验法进行检验。本例中各点重复测试次数的最小值为 4。故可直接用探索性分析中的 Levine 中位数检验法进行检验。

在 SPSS 中具体解题步骤如下:

- ① 在 SPSS 数据编辑窗口打开数据文件 data10-10.sav。
- ② 按 Analyze→Descriptive Statistics→Explore 顺序, 打开 Explore 对话框(见图 2-99)。
- ③ 在左边的变量名源框中, 选中抗拉强度、延伸率变量, 通过中间的右移箭头将它们移到 Dependent List 框中, 用同样的方法将变量含碳量移到 Factor List 框中。在 Display

选项中，选择 Plots，单击 Plots 按钮，打开 Plots 对话框，见图 2-104。

④ 在 Plots 对话框中的 Boxplots 选择项中，选择 None，不输出箱图。在 Spread vs Level with Lenvene Test 选择项中，选择 Untransformed 选项。要求做原始数据状态下的方差齐性检验。

单击 Continue 按钮，返回 Explore 对话框。

⑤ 单击 OK 按钮运行，在输出窗口中得到如表 10-55 所示的中位数方差齐性检验结果。

表 10-55 中位数方差齐性检验

Test of Homogeneity of Variance					
		Levene Statistic	df1	df2	Sig.
抗压强度	Based on Mean	3.896	9	58	.001
	Based on Median	3.297	9	58	.003
	Based on Median and with adjusted df	3.297	9	25.091	.009
	Based on trimmed mean	3.854	9	58	.001
延伸率	Based on Mean	1.310	9	58	.252
	Based on Median	.918	9	58	.516
	Based on Median and with adjusted df	.918	9	43.498	.519
	Based on trimmed mean	1.288	9	58	.263

⑥ 结果与结论

从表 10-55 抗压强度的第二行的 Based on Median 中可以看到，抗压强度变量的 L 值为 3.297，第一自由度为 9 ($m-1=10-1=9$)，第二自由度为 58 ($n-m=68-10=58$)，在自变量各点的观测值的方差相等的原假设下，出现目前统计量的值或更极端值的概率为 0.001，小于 0.05，故拒绝原假设，而认为至少有两个点上的抗压强度观测值的方差是不想等的。

同样，从表 10-55 延伸率的第二行的 Based on Median 中可以看到，延伸率变量的 L 值为 0.918，第一自由度为 9，第二自由度为 58，在自变量各点的观测值的方差相等的原假设下，出现目前统计量的值或更极端值的概率为 0.516，大于 0.05，故现有证据不支持拒绝原假设，而认为在各点上的延伸率观测值的方差是相等的。

3. 无重复试验下残差诊断

当没有重复测量数据时，对模型各个假定条件的检验就只有依靠残差来进行了。由于用残差检验模型假定条件的过程中，许多方法不太容易给出简明的理论解释，因此，检查如同大夫切脉一样，需要由表及里，故称这些方法为残差诊断。残差诊断的目的就是要寻找残差中表明问题的现象，从而为改进模型来提供有用的信息。

如前所述，所谓残差就是观测值与拟合值（预测值）之差，用符号表示为 $\hat{e}_i = y - \hat{y}_i$ ($i=1,2,\dots,n$)。它是回归方程未能解释的量。在模型正确时，残差就是观测误差的估计。在这种条件下，残差具有如下性质

$$E(\hat{e}_i) = 0$$

$$\text{Var}(\hat{e}_i) = \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{L_{xx}} \right] \sigma^2$$

$$\text{Cov}(\hat{e}_i, \hat{e}_j) = - \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{L_{xx}} \right] \sigma^2$$

式中, $i \neq j$, $i, j = 1, 2, \dots, n$ 。

在进行诊断时, 有时要用到如下的标准化残差 (也称学生氏残差)

$$r_i = \frac{\hat{e}_i}{s \sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{L_{xx}}}} \quad i = 1, 2, \dots, n$$

在模型的4个条件都满足且 n 较大时, r_1, r_2, \dots, r_n 可近似视为来自于 $N(0,1)$ 的独立样本。

(1) 模型误差的诊断

残差诊断的基本工具是各种各样的残差图。各种残差图中较为常用的是 $\hat{y} - \hat{e}$ 图 (以 \hat{y} 为横轴, 残差或学生化残差为纵轴的散点图) 和 $x - \hat{e}$ 图 (以 x 为横轴, 残差或学生化残差为纵轴的散点图)。模型系统误差的诊断主要看残差图中有无非线性的趋势。

$\hat{y} - \hat{e}$ 图 ($x - \hat{e}$ 图) 的识别方法如下:

① 在模型条件1-3满足的条件下, $\hat{y} - \hat{e}$ 图典型的特征是: 散点构成的点云呈水平带状、带宽一致、关于横轴对称、绝大部分点应落在 $e = \pm 2$ 两条平行线之间、散点杂乱出现、不呈前后点之间上下位置的规律性。

② 当模型的条件1可能不满足, 即可能不全是随机误差, 而存在系统误差时, 其散点图的症状是点云呈宽度一致的带状, 但不是水平的, 而是呈现出某种趋势, 或者点云关于横轴不对称。如点云呈倒钟形、S形等。

例 10.12 以例 10.9 中的腐蚀深度 y 与腐蚀时间 x 的数据为例 (数据存放在 data10-08.sav 中), 试用 $\hat{y} - \hat{e}$ 图对其模型中是否存在系统误差进行诊断。

在 SPSS 中的具体解题步骤如下:

① 在 SPSS 的数据编辑窗口中, 打开 data10-08.sav。

② 按 Analyze → Regression → Linear 顺序展开 Linear Regression 对话框, 见图 10-10。在左侧变量名源框中, 选择腐蚀深度变量, 将其移入到 Dependent 框中, 作为因变量, 选择腐蚀时间变量, 将其移入到 Independent(s) 框中, 作为自变量。

③ 单击 Plots 按钮, 打开 Linear Regression: Plots 对话框, 见图 10-13, 在左侧变量框中从上到下依次列出的是:

- DEPENDENT: 因变量。
- PRED: 标准化预测值。
- RESID: 标准化残差。

- RESID: 剔除残差。
- DJPRED: 校正后的预测值。
- RESID: 学生化残差。
- DRESID: 学生化剔除残差。

它们可用来作为Y和X轴的变量。

在本例中，选择 ZPRED 选项将其移入到 X 轴的框中，作为 X 轴变量，选择 SRESID 选项将其移入到 Y 轴的框中，作为 Y 轴变量。其他保持系统默认选择。

单击 Continue，返回到 Linear Regression 对话框。

④ 单击 OK 按钮运行，除同例 10.8 一样，可在输出窗口中得到几张计算结果表外，另外还得到一张本例中所需要的散点图，见图 10-14。

为使读者在图中看得更加清晰，本例在输出窗中，仿例10-1中的做法，在图形编辑器中，对图进行了修改，加上了拟合曲线。

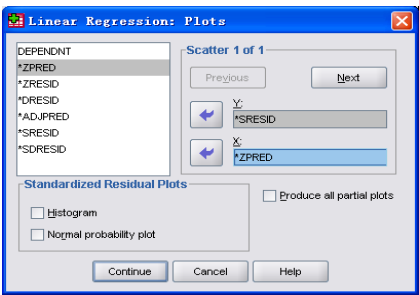


图 10-13 Linear Regression: Plots 对话框

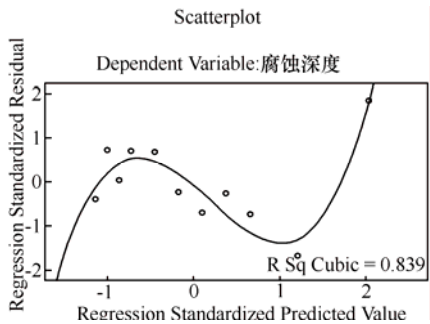


图 10-14 标准化预测值与学生化残差散点图

⑤ 诊断结果

从图10-14可见，点云图为S形，说明残差图中有非线性的趋势存在，即模型1的条件（误差是随机的条件）并不满足，表明误差中还可能存在系统误差，所以，用线性拟合得到的回归方程虽然是显著的，但它并不是最佳拟合方程，可以改用非线性的回归加以拟合可能效果会更好。

（2）独立性诊断

对独立性假定的诊断，在许多实际问题中，是可以用其实际的背景和依据专业知识来加以判定。当测量顺序不影响变量之间的关系时，可以认为它是独立的。例如，在例 10.1 中，改变弹簧伸长长度的顺序时，测定的质量同测定的先后顺序是无关的，而只是与弹簧伸长的长度有关，因此，它是独立的。

在较为特殊而又常见的时间序列（观察点是按时间轴取得的）数据中，可以通过构造残差对时间图来观察残差是否存在序列相关。在时间上相近的残差倾向于有相同符号时，表明时间序列中存在正序列相关，当相近的残差符号交替不同时，表明序列存在负

相关。而出现与这两种情况不同的情形，则表明序列无明显相关性。

序列的相关性也可以用 Durbin-Waston 统计量

$$d = \frac{\sum_{i=1}^{n-1} (\hat{e}_{i+1} - \hat{e}_i)^2}{\sum_{i=1}^n \hat{e}_i^2}$$

来检验。其中， \hat{e}_i 为时刻 i 时的残差， n 为时间点的总数。

检验一阶序列相关的原假设为 H_0 ：所有值之间不相关，被择假设 H_1 ：现在的值与前一时刻的值相关，且相关系数为常数。

对于 Durbin-Waston 统计量 d 值，当它接近于 2.0 时，表明序列不相关，当 $d < 2.0$ 时，表明存在正序列相关，当 $d > 2.0$ 时，表明存在负序列相关。当 $d < d_{L\alpha}$ 或 $d > 4 - d_{L\alpha}$ 时，认为有序列相关。J.Durbin 和 G.S.Watson 在 1951 年做出了 d 的临界值表。实际应用中当 d 小于 1.5 或大于 2.5 时，怀疑该时间序列中存在正相关或负相关。

例 10.13 某公司 1986~1997 年间各季度某商品的销售量数据已存放在数据文件 data10-11 中，用残差对时间的散点图来观察残差是否存在序列相关。

① 在 SPSS 的数据编辑窗口中，打开 data10-11.sav。

② 按 Analyze→Regression→Linear 顺序展开 Linear Regression 对话框，见图 10-10。在左侧变量名源框中，选择销售量变量，将其移入到 Dependent 框中，作为因变量，选择时间点变量，将其移入到 Independent(s) 框中，作为自变量。

③ 单击 Save 按钮，打开 Linear Regression: Save 对话框，见图 10-15。

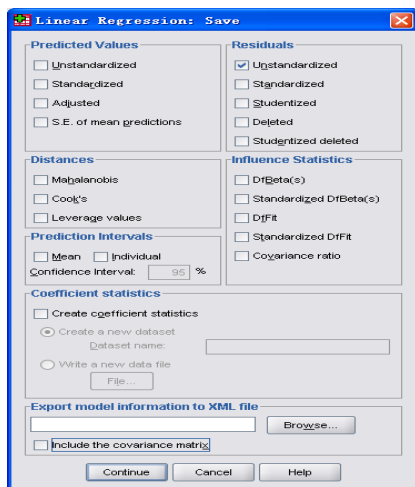


图 10-15 Linear Regression: Save 对话框

在 Residuals 项下，选择 Unstandardized 选项，要求输出残差于工作的数据文件中。

关闭 Include the covariance matrix 选项，即输出中不包括协方差矩阵。

单击 Continue，返回到 Linear Regression 对话框。

④ 单击 OK 按钮运行，除同例 10.8 一样，可在输出窗口中得到几张计算结果表外，在工作数据文件中，增加了一列存放预测误差的新变量 RES_1 及其值。将存放预测误差的新变量 RES_1 及其值的工作数据文件另存为 data10-11a.sav。

现在工作的数据文件为 data10-11a.sav。

按例 10.1 中第一步制作散点图的方法，可

得残差与时间点的散点图，见图 10-16，及销售量与时间点的趋势图，见图 10-17。

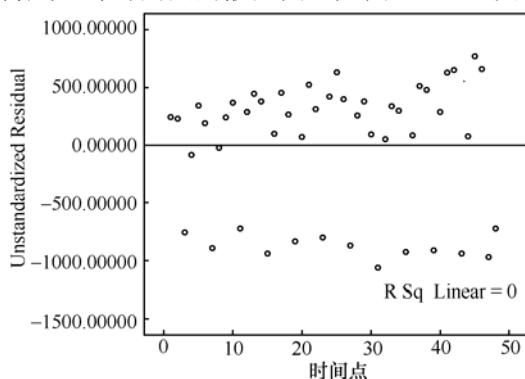


图 10-16 残差与时间点的散点图

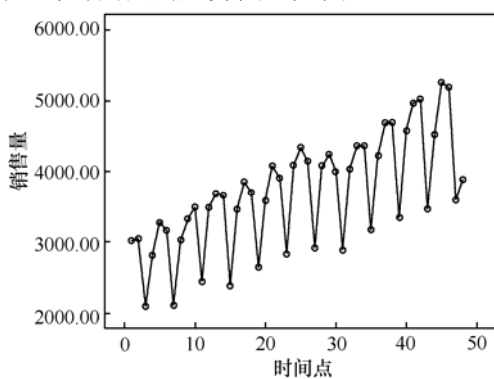


图 10-17 销售量与时间点的趋势图

从图10-16中可见，散点图中的点云呈不规律排列，表明时间序列中不存在一阶序列相关。

现在用残差 RES_1 和上面给出的 Durbin-Waston 统计量 d 的计算公式，可以计算统计量 d 的值，具体过程如下：

按 Transform→Create Time Series 顺序，打开 Create Time Series 对话框，见图 10-18。将左侧变量名框中的 RES_1 变量移到 New Variable(s)框中。在 Function 下拉式选项中，选择 Lag 选项，用来计算 $\hat{e}_{i+1} - \hat{e}_i$ 的结果，其他保持系统默认选项。

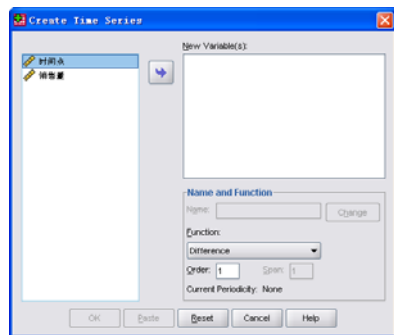


图 10-18 Create Time Series 对话框

单击 OK 按钮运行，则在工作数据文件中出现一行存放 $\hat{e}_{i+1} - \hat{e}_i$ 结果的新变量 RES_1_1。

然后，用在第 2 章中提到的建立新变量（派生指标）的方法，在工作数据文件中，生成新变量 A（用来存放计算得到的残差 RES_1 的平方值）和新变量 B（用来存放计算得到的 RES_1_1 的平方值），再用描述统计的程序，在输出窗口中得到 A、B 和的计算结果，见表 10-56。

表 10-56 计算结果

Descriptive Statistics		
	N	Sum
A	48	1.52E7
B	47	3.27E7
Valid N (listwise)	47	

所以，所要计算的 $d = \frac{B}{A} = \frac{3.27 \times 10^7}{1.52 \times 10^7} = 2.15$ ，由于它

接近于2.0，也同样支持以上的诊断。说明极可能满足独立性条件。

但即使满足独立性这一条件，也并不是说，就可以用线性回归模型最佳地来描述时间和销售量之间的关系了。

从图 10-17 中可见,销售量和时间之间明显存在季节效应。用时间序列季节效应模型要远比线性回归模型来得恰当。

(3) 方差齐性的诊断

在条件 3 满足时, y 的方差与 x 无关, 即随机误差的变异性与 x 无关, 在 $\hat{y} - \hat{e}$ 图上, 反映方差齐性的特点是点云呈带状, 沿横轴水平分布, 且带状的宽度不随 x 的变化而变化。否则, 如出现喇叭状、腰鼓状等宽度不等的情况时, 表明误差的变异性可能与 x 有关, 它不是常数, 也即表明数据不满足方差齐性的假定。

方差不齐, 对点预测影响不大, 但对区间估计和区间预测及假设检验的精度影响很大。此时, 可以尝试用加权回归或方差稳定化变换等方法来进行处理。

例 10.14 仍以例 10.13 中时间—销售量数据为例, 数据存放在 data10-11.sav 中, 对其方差齐性进行诊断。

按例 10.12 中的做法, 可得学生化残差与回归标准预测值的散点图, 见图 10-19。

从图 10-19 可见, 点云的整体形态呈水平, 未表明“模型是线性的”有问题, 同样, 由于点云带宽不随 x 的变化而变化, 因而, 没有证据支持方差不齐的说法, 但用线性模型来描述销售量和时间之间的关系同例 10.13 中给出的结论相同, 依然是不合适的。图中给出的用线性模型拟合时的 $R^2 = 4.633E-7$, 接近于 0, 表明用时间作为一元回归的自变量, 几乎对因变量销售量不起任何解释作用。同样说明了一个问题, 即用残差图来诊断时, 还要结合因变量和自变量的散点图来一起分析。

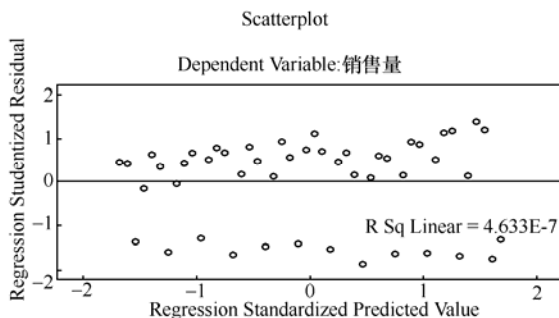


图 10-19 学生化残差与回归标准预测值的散点图

(4) 正态假定的诊断

对因变量 Y 做正态性检验在前面已多次提到, 它是与残差图形正态假定诊断是等价的一个方法。在这里主要介绍, 用残差的图形来对正态假定进行诊断。

在不重复测试情况下, 对误差正态分布的检验可用学生化残差作为样本, 用第 2 章 2.4 节提到的探索性分析方法来做 Q-Q 图和 Kolmogorov 修正检验。也可在 Linear Regression: Plots 对话框 (见图 10-13) 中, 选择 Normal probability plot 选项, 利用输出的残差正态概率图, 来检验残差的正态性。

例 10.15 对例 10.9 中的腐蚀深度 y 与腐蚀时间 x (数据存放在 data10-08.sav 中) 建立一元回归方程中所产生的残差的正态性进行诊断。

在 SPSS 中的具体解题步骤如下:

1. 在 SPSS 的数据编辑窗口中, 打开 data10-08.sav。

2. 按 Analyze→Regression→Linear 顺序展开 Linear Regression 对话框, 见图 10-10。在左侧变量名源框中, 选择腐蚀深度变量, 将其移入到 Dependent 框中, 作为因变量, 选择腐蚀时间变量, 将其移入到 Independent(s)框中, 作为自变量。

3. 单击 Plots 按钮, 打开 Linear Regression: Plots 对话框, 见图 10-13。

选择 Normal Probability plot 选项, 要求输出标准化残差的正态概率图。

单击 Continue 按钮, 返回到 Linear Regression 对话框。

4. 按 OK 按钮运行, 除同例 10.8 一样, 可在输出窗口中得到几张计算结果表外, 另外还得到一张本例中所需要的残差的概率图, 见图 10-20。

5. 诊断结果

从图 10-20 可见, 标准化残差的各点的整体基本都在图中对角线上或其周围, 由此可以认为与正态性假设无明显相悖。

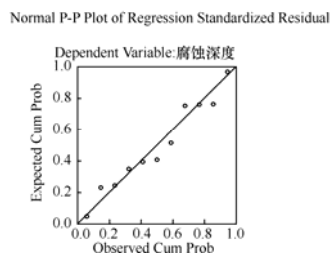


图 10-20 标准化残差的正态概率图

10.5.3 曲线估计-一元非线性回归

在实际研究的问题中, 通过变量间观测数据所作的散点图, 可以观察到变量之间的关系并非都呈直线, 许多散点图明显地显示出曲线的特征。因此, 要使回归结果能客观地反映实际问题中变量的非线性关系, 就需要进行非线性回归分析, 即曲线回归。

10.5.3.1 一元非线性回归模型的一般求法

一元非线性回归的一般模型为: $Y = f(X, \theta) + \varepsilon$, 其中未知参数 $\theta = (\beta_0, \beta_1)'$, 它的主要标志是回归方程中的未知参数不都是以线性的形式出现在方程中, 但大部分是通过某种方法的线性转换 (如倒数变换、开根号变换、对数变换、指数变换、幂函数变换等) 将其线性化的, 对于线性化后的方程可用一元线性回归方程中介绍的 LS 法 (最小二乘法) 估计其参数, 再还原成原来的曲线形式, 这就是非线性回归模型的一般求法。

常见一元非线性回归模型及线性化方法:

1. 多项式模型

(1) 模型

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n$$

(2) 线性化方法

令 $x_1 = x$ 、 $x_2 = x^2$ 、 \cdots 、 $x_n = x^n$ ，可将多项式模型转换为多元线性回归方程

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n,$$

再利用最小二乘法确定系数 β_0 、 β_1 、 \cdots 、 β_n （下一节介绍），代入原方程即可。

2. 复合曲线模型

(1) 模型

$$y = \beta_0 \beta_1^x$$

(2) 线性化方法

令 $y' = \ln y$ ， $\beta'_0 = \ln \beta_0$ ， $\beta'_1 = \ln \beta_1$ ，可将复合曲线模型转换为线性回归方程

$$y' = \beta'_0 + \beta'_1 x,$$

在用 LS 法求得 β'_0 和 β'_1 后，回代等式中可得 β_0 和 β_1 ，再代入原方程即可。

3. 生长曲线模型

(1) 模型

$$y = e^{(\beta_0 + \beta_1 x)}$$

(2) 线性化方法

令 $y' = \ln y$ ，则可将生长曲线模型转换为线性回归方程

$$y' = \beta_0 + \beta_1 x$$

在用 LS 法求得 β_0 和 β_1 后，再代入原方程即可。

4. 对数函数模型

(1) 模型

$$y = \beta_0 + \beta_1 \ln x$$

(2) 线性化方法

令 $x' = \ln x$ ，则可将对数函数模型转换为线性回归方程

$$y = \beta_0 + \beta_1 x'$$

在用 LS 法求得 β_0 和 β_1 后，再代入原方程即可。

5. S 形曲线模型

(1) 模型

$$y = \frac{1}{\beta_0 + \beta_1 e^{-x}}$$

(2) 线性化方法

令 $y' = \frac{1}{y}$ ， $x' = e^{-x}$ ，则可将 S 形曲线模型转换为线性回归方程

$$y' = \beta_0 + \beta_1 x'$$

在用 LS 法求得 β_0 和 β_1 后, 再代入原方程即可。

6. 指数函数模型

(1) 模型

$$y = \beta_0 e^{\beta_1 x}$$

(2) 线性化方法

令 $y' = \ln y$, $\beta_0' = \ln \beta_0$, 则可将指数函数模型转换为线性回归方程

$$y' = \beta_0' + \beta_1 x$$

在用 LS 法求得 β_0' 和 β_1 后, 回代得 β_0 , 再将 β_0 和 β_1 代入原方程即可。

7. 倒幂函数模型

(1) 模型

$$y = \beta_0 + \beta_1 \frac{1}{x}$$

(2) 线性化方法

令 $x' = \frac{1}{x}$, 则可将逆函数模型转换为线性回归方程

$$y = \beta_0 + \beta_1 x'$$

在用 LS 法求得 β_0 和 β_1 后, 再代入原方程即可。

8. 幂函数模型

(1) 模型

$$y = \beta_0 x^{\beta_1}$$

(2) 线性化方法

令 $y' = \ln y$, $\beta_0' = \ln \beta_0$, $x' = \ln x$, 则可将幂函数模型转换为线性回归方程

$$y' = \beta_0' + \beta_1 x'$$

在用 LS 法求得 β_0' 和 β_1 后, 回代得 β_0 , 再将 β_0 和 β_1 代入原方程即可。

9. 双曲线模型

(1) 模型

$$\frac{1}{y} = \beta_0 + \beta_1 \frac{1}{x}$$

(2) 线性化方法

令 $y' = \frac{1}{y}$, $x' = \frac{1}{x}$, 则可将双曲线模型转换为线性回归方程

$$y' = \beta_0 + \beta_1 x'$$

在用 LS 法求得 β_0 和 β_1 后, 再代入原方程即可。

10. Logistic 曲线模型

(1) 模型

$$y = \frac{a}{1 + \beta_0 e^{-\beta_1 x}}$$

其中, a 为任意一个由使用者指定的大于 0 的数。

(2) 线性化方法

将原方程变形得, $\frac{a-y}{y} = \beta_0 e^{-\beta_1 x}$, 令 $y' = \ln \frac{a-y}{y}$, $\beta'_0 = \ln \beta_0$, $\beta'_1 = -\beta_1$, 则可将 Logistic 曲线模型转换为线性回归方程

$$y' = \beta'_0 + \beta'_1 x$$

在用 LS 法求得 β'_0 和 β'_1 后, 回代等式中可得 β_0 和 β_1 , 再代入原方程即可。

10.5.3.2 实例分析

1. 寻找相对最优的曲线拟合方程

例 10.16 出钢时所用盛钢水的钢包, 由于钢水对耐火材料的侵蚀, 容积会不断扩大, 使用次数与增大容积之间的观测值见表 10-57, 数据已存放在 data10-12.sav 中, 试求使用次数 x 与增大容积 y 之间的回归函数。

表 10-57 使用次数与增大容积记录表

使用次数	增大容积	使用次数	增大容积	使用次数	增大容积	使用次数	增大容积
2	6.42	6	9.70	10	10.49	14	10.60
3	8.20	7	10.00	11	10.59	15	10.80
4	9.58	8	9.93	12	10.60	16	10.60
5	9.50	9	9.99	13	10.80		

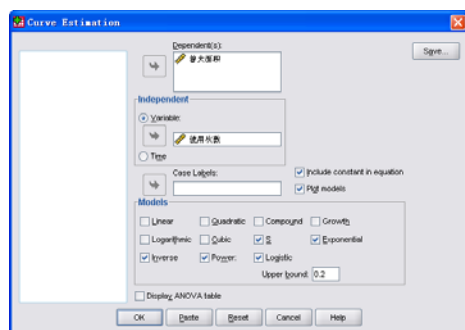


图 10-21 Curve Estimation 对话框

在 SPSS 中的具体解题步骤如下:

① 在 SPSS 的数据编辑窗口中, 打开 data10-12.sav。

② 按 Analyze → Regression → Curve Estimation 顺序展开 Curve Estimation 对话框, 见图 10-21。在左侧变量名源框中, 选择增大容积变量, 将其移入到 Dependent 框中, 作为因变量, 选择使用次数变量, 将其移入到 Independent(s)框中, 作为自变量。

在 Models 选项中, 有 11 个选项, 它们分

别是：

- Linear：一元线性模型，同式（10.1）。
- Quadratic：二次函数，为 $n=2$ 时的多项式模型。
- Compound：复合函数，见复合曲线模型。
- Growth：生长函数，见生长曲线模型。
- Logarithmic：对数函数，见对数函数模型。
- Cubic：三次函数，为 $n=3$ 时的多项式模型。
- S：S 形曲线，见 S 形曲线模型。
- Exponential：指数函数，见指数函数模型。
- Inverse：倒数函数，见倒幂函数模型。
- Power：幂函数，见幂函数模型。
- Logistic：逻辑斯蒂函数，见 Logistic 曲线模型。

由于本题未对原始数据制作散点图，因此，在无法确定使用次数和增大面积之间究竟呈线性关系还是呈何种曲线关系时，可以用探究性的方法，选中全部11个选项来加以比较研究。

由于选择了Logistic选项，因此需要在Upper bound后框中输入一个大于或等于因变量的最大值的数作为Logistic曲线模型中 a 的上界值。否则，Logistic选项将不被执行。

其他保持系统默认选择。

③ 单击 OK 按钮运行，则在输出窗口中得到 4 张表和一个图，分别为表 10-58、表 10-59、表 10-60、表 10-61 和图 10-22。

表 10-58 模型描述

Model Description		
Model Name	1	MOD_7
Dependent Variable	1	增大面积
Equation	2	Linear
	3	Logarithmic
	4	Inverse
	5	Quadratic
	6	Cubic
	7	Compound ^a
	8	Power ^a
	9	S ^a
	10	Growth ^a
	11	Exponential ^a
		Logistic ^{a, b}
Independent Variable		使用次数
Constant		Included
Variable Whose Values Label Observations in Plots		Unspecified
Tolerance for Entering Terms in Equations		.0001

a. The model requires all non-missing values to be positive.
b. For all dependent variables, the theoretical upper bound is set to 11.

表 10-59 样品处理摘要

Case Processing Summary	
	N
Total Cases	15
Excluded Cases ^a	0
Forecasted Cases	0
Newly Created Cases	0

a. Cases with a missing value in any variable are excluded from the analysis.

④ 结果与讨论

表 10-58 给出了模型的描述，从上到下依次是：

模型名称：MOD_7，因变量 1：增大面积，方程 1：一元线性回归方程，方程 2：对数函数，方程 3：倒幂函数，方程 4：二次函数，方程 5：三次函数，方程 6：复合

函数, 方程 7: 幂函数, 方程 8: S 形曲线, 方程 9: 生长函数, 方程 10: 指数函数, 方程 11: 逻辑斯蒂函数, 自变量: 使用次数, 常量: 包含, 观测值的值标签出现在图中的变量: 没有指定, 方程中输入项的容忍度为: 0.0001。

表 10-60 变量处理摘要

		Variables	
		Dependent	Independent
		增大面积	使用次数
Number of Positive Values		15	15
Number of Zeros		0	0
Number of Negative Values		0	0
Number of Missing Values	User-Missing	0	0
	System-Missing	0	0

表 10-61 模型摘要和参数估计

Dependent Variable: 增大面积									
Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.661	25.306	1	13	.000	7.938	.213		
Logarithmic	.857	78.012	1	13	.000	6.290	1.742		
Inverse	.967	376.929	1	13	.000	11.358	-9.482		
Quadratic	.862	37.348	2	12	.000	6.031	.763	-.031	
Cubic	.916	40.058	3	11	.000	4.199	1.660	-.146	.004
Compound	.599	19.411	1	13	.001	7.903	1.024		
Power	.810	55.509	1	13	.000	6.527	.198		
S	.956	283.465	1	13	.000	2.454	-1.100		
Growth	.599	19.411	1	13	.001	2.067	.024		
Exponential	.599	19.411	1	13	.001	7.903	.024		
Logistic	.860	79.566	1	13	.000	.049	.807		

The independent variable is 使用次数.

表 10-59 显示的是样品处理摘要, 从上到下依次为:

全部样品: 15, 剔除样品: 0, 预测样品: 0, 最新建立的样品: 0。

表 10-60 显示了变量处理的摘要, 从上到下依次为:

正值数量: 因变量增大面积为 15 个, 自变量使用次数为 15 个, 0 的数量: 因变量增大面积为 0 个, 自变量使用次数为 0 个, 负值数量: 因变量增大面积为 0 个, 自变量使用次数为 0 个, 缺失值数量: 用户缺失值 (因变量增大面积为 0 个, 自变量使用次数为 0 个)、系统缺失值 (因变量增大面积为 0 个, 自变量使用次数为 0 个)。

表 10-61 显示的是模型摘要和参数估计, 第一列列出的是方程的名称, 第二列为 R^2 , 第三列列出的是方差分析的 F 值, 第四列列出的是 F 检验的第一自由度 df_1 , 第五列列出的是 F 检验的第二自由度 df_2 , 第六列为 F 检验的显著性水平, 即 P 值, 第七列列出的是方程中的常量, 第八列至第十列列出的是回归系数 b_1 、 b_2 和 b_3 的值。

从表 10-61 中可见, 全部回归方程均有统计上的极显著性意义 ($P < 0.01$), 从 R^2 上来看, 相对而言, 倒幂函数模型的拟合效果最好, 其 R^2 最大为 0.967。

而图 10-22 显示了 11 种方程的拟合曲线图, 由于它们全部重叠地出现在一张图上, 因此看起来不是那么清晰。

因此, 有必要在上述探索性分析的基础上, 在只选 Inverse 选项的基础上重做一遍, 这样, 可以在输出窗中得到图 10-23。从中可见, 所有点均最近地出现在倒幂函数的曲线上或周围, 因此, 用倒幂函数来做曲线拟合是恰当的。

结合倒幂函数的公式和表 10-61 中的模型参数的估计, 可知, 所要求的使用次数 x 与增大容积 y 之间的回归函数为

$$y = 11.358 - \frac{9.482}{x}$$

回归方程具有统计上的极显著性意义 ($P=0.000<0.01$)。

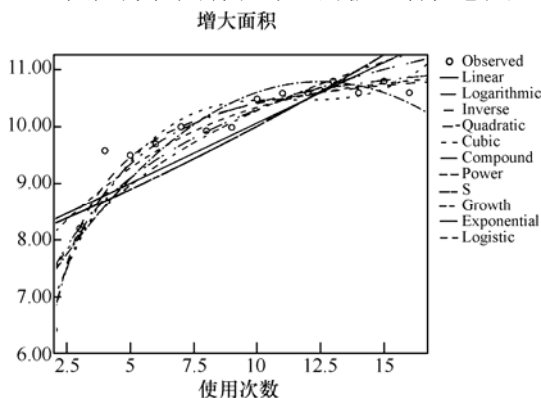


图 10-22 拟合曲线图

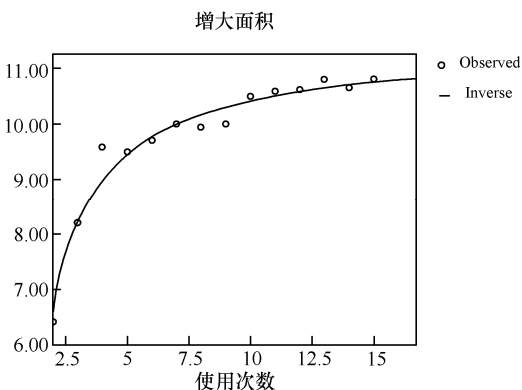


图 10-23 倒幂函数拟合曲线图

2. 对时间序列进行曲线估计

例 10.17 从《中国统计年鉴 2007》(中国统计出版社, 2007 年)中得到 1978 年~2006 年间社会消费品零售总额(单位: 亿元)的数据资料, 数据已存放在 data10-13.sav 中, 试对它们之间的关系进行曲线回归分析, 并对 1978~2010 年的社会消费品零售总额进行预测。

在 SPSS 中的具体解题步骤如下:

- ① 在 SPSS 的数据编辑窗口中, 打开 data10-13.sav。
- ② 作社会消费品零售总额与年份的散点图。

我们可以用例 10.1 中介绍的方法, 将年份作为横坐标, 总额作为纵坐标, 在直角坐标系描出各点, 见图 10-24, 从图中可见, 这些点呈一条曲线。根据曲线估计中提供的 11 个函数的曲线特征, 可以用来拟合上述图形的函数有: 幂函数、指数函数、二次函数和三次函数。

③ 按 Analyze→Regression→Curve Estimation 顺序展开 Curve Estimation 对话框, 见图 10-21。在左侧变量名源框中, 选择 **社会消费品零售总额** 变量, 将其移入到 **Dependent** 框中, 作为因变量, 选择 **年份** 变量, 将其移入到 **Independent(s)** 框中, 作为自变量。选择 **Time** 选项。

在 **Models** 选项中, 选择 **Quadratic**、**Cubic**、**Power**、**Exponential** 选项, 分别用二次函数、三次函数、幂函数和指数函数对数据资料进行曲线估计, 从中选优。

其他保持系统默认选择。

④ 单击 **OK** 按钮运行, 则同例 10.16 一样在输出窗口中得到 4 张表和一个图, 其中对分析有用的是第四张表和最后一个图。见表 10-62 和图 10-25。

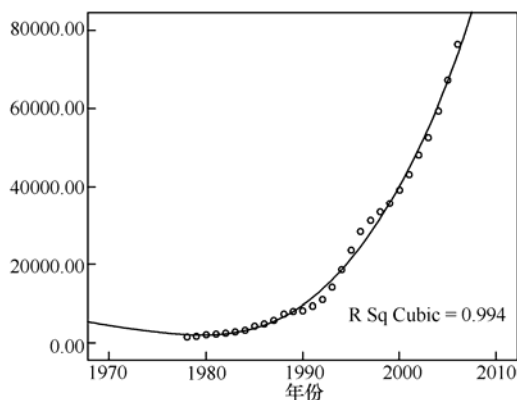


图 10-24 社会消费品零售总额与年份的散点图

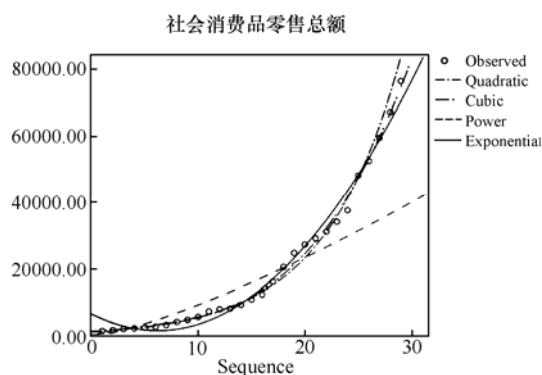


图 10-25 四种拟合曲线图

表 10-62 模型摘要和参数估计

Model Summary and Parameter Estimates									
Dependent Variable: 社会消费品零售总额									
Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Quadratic	.987	953.866	2	26	.000	6.756E3	-1.639E3	132.133	
Cubic	.995	1.816E3	3	25	.000	230.765	768.904	-65.200	4.385
Power	.856	160.241	1	27	.000	446.258	1.322		
Exponential	.995	5.654E3	1	27	.000	1.368E3	.142		

从表 10-62 中可见,所选 4 个函数模型所作的曲线回归方程均有统计上的极显著性意义,全部 F 检验的 P 值为 0.000,均小于 0.01,而三次函数和指数函数对应的 R^2 最大为 0.995,二次函数的 R^2 次之为 0.987,从图 10-25 中同样可以见到,所有点最靠近三次和指数函数拟合曲线。因此,可用三次函数和指数函数来进行拟合。本例选用三次函数来做曲线回归。

三次函数的回归方程为

$$\hat{y} = 230.765 + 768.904x - 65.2x^2 + 4.385x^3$$

⑤ 预测

按 Analyze→Regression→Curve Estimation 顺序展开 Curve Estimation 对话框,见图 10-21。在左侧变量名源框中,选择社会消费品零售总额变量,将其移入到 Dependent 框中,作为因变量,选择年份变量,将其移入到 Independent(s) 框中,作为自变量。选择 Time 选项。

在 Models 选项中,选择 Cubic 选项。

关闭其他系统默认选项。

单击 Save 按钮,打开 Curve Estimation: Save 对话框,见图 10-26。

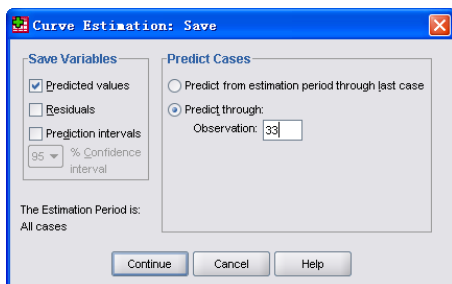


图 10-26 Curve Estimation: Save 对话框

在 Save Variables 项中选择 Predicted Values 选项, 表示要在工作的数据文件中存取预测值, 在 Predict Cases 项中选择 Predict through 选项, 并在 Observation 框中输入 33, 表示对原有 29 年观测值作出预测外, 另外还向外预测 4 年的值, 即预测到 2010 年。

单击 Continue 返回到 Curve Estimation 对话框。

单击 OK 运行, 则在工作数据文件中, 增加一列新变量, 它是用上面二次函数的回归方程得到的从 1978 年直到 2010 年的每年社会消费品零售总额的预测值。据此, 可以得到 2010 年的社会消费品零售总额的预测值为 112191.79754 亿元, 并将该数据文件存放在 data10-13a.sav 中。

3. 对其他非线性模型进行曲线估计

例 10.18 测定某水稻土用 HCl 和 Na₂CO₃ 调酸后的 pH 和铵态 N 含量的结果, 见表 10-63, 数据存放在 data10-14.sav 中, 试对其做 4 次多项式回归。

表 10-63 某水稻土用 HCl 和 Na₂CO₃ 调酸后的 pH 和铵态 N 含量的测定结果

pH(x)	2	3	4	5	6	7	8	9
铵态 N(y)	13.0	9.2	6.6	4.7	4.0	7.1	13.2	20.0

【题析】四次多项式模型为

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$$

这在曲线估计程序中, 没有现成的选项可以来完成这个模型的拟合, 它可以通过数据转换后, 在线性回归程序中来处理, 也可以用以下的方式在非线形回归程序中来实现。具体做法如下:

① 在 SPSS 的数据编辑窗口中, 打开 data10-14.sav。

② 按 Analyze→Regression→Nonlinear Regressions 顺序展开 Nonlinear Regressions 对话框, 见图 10-27。在左侧变量名源框中, 选择铵态 N 变量, 将其移入到 Dependent 框中, 作为因变量, 在 Model Expression 框中, 输入四次多项式的回归模型

$$a+b*\text{pH}+c*\text{pH}^2+d*\text{pH}^3+e*\text{pH}^4$$

单击 Parameters 按钮, 弹出 Nonlinear Regressions: Parameters 对话框, 见图 10-28。在 Name 框中输入上面模型中用到的参数名 a, 在 Starting Value 框中, 输入 1 (也可以是其他的数字), 对其进行初始值定义, 单击 Add 按钮, 将定义好的参数添加到参数名框中。循环做参数初始值的定义, 直到将最后一个参数 e 定义好为止。

单击 Continue 按钮返回到 Nonlinear Regressions 对话框。

单击 Options 按钮, 弹出 Nonlinear Regressions: Options 对话框, 见图 10-29。

选择 Bootstrap estimates of standard error 选项, 要求使用 Bootstrap 抽样方法估计参数的标准误。

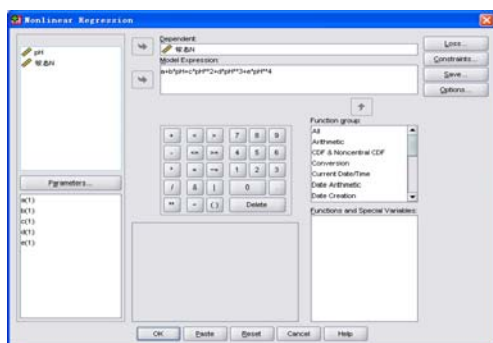


图 10-27 Nonlinear Regressions 对话框

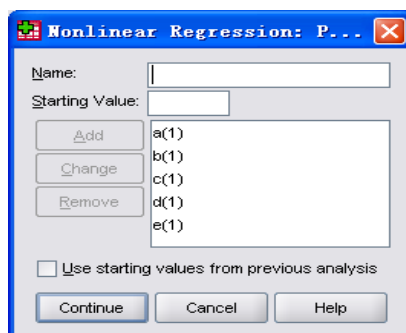


图 10-28 Nonlinear Regressions Parameters 对话框

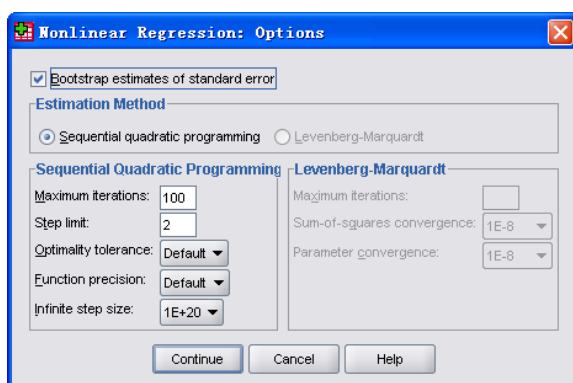


图 10-29 Nonlinear Regressions: Options 对话框

Bootstrap 方法由 Efron 于 1979 年提出,它是基于大量计算的一种模拟抽样统计推断方法。其基本思想为:在原始数据的范围内做有放回的抽样,样本量为 n ,原始数据中每个观察单位每次抽到的概率都为 $1/n$,所得样本被称为 Bootstrap 样本,于是可得参数 θ 的一个估计值,若重复抽取这样的样本 1000 次,则得到该参数 1000 个估计值,以此可以求得 θ 的一些性质。这样可以用它来判断参数估计值是否准确,计算出更准确的置信区间,判断得出的统计学结论是否正确。

在 Maximum iterations 框中输入 100,规定最大迭代次数为 100。其他采用系统默认选项,单击 Continue 按钮返回到 Nonlinear Regressions 对话框。

单击 OK 按钮,则在输出窗口中,得到 4 张计算结果表,见表 10-64、表 10-65、表 10-66 和表 10-67。

③ 结果与讨论

表 10-64 显示了迭代过程记录,模型中 5 个参数的迭代初始值都被设置为 1,在迭代到 13 次时,模型达到收敛标准,也即找到了最优解。

表 10-64 迭代过程

Iteration History ^b						
Iteration Number ^a	Residual Sum of Squares	Parameter				
		a	b	c	d	e
0.1	8.691E7	1.000	1.000	1.000	1.000	1.000
1.1	20263.038	1.000	.998	.984	.867	-.116
2.1	105.071	.985	.929	.672	-.223	.018
3.1	73.136	1.272	1.449	1.322	-.446	.034
4.1	13.678	5.777	5.298	-1.878	.159	.001
5.1	.995	9.036	8.071	-4.186	.595	-.024
6.1	.987	9.940	7.238	-3.929	.563	-.022
7.1	.987	9.941	7.231	-3.925	.562	-.022
8.1	.987	10.497	6.723	-3.768	.542	-.021
9.1	.987	10.313	6.890	-3.820	.549	-.022
10.1	.987	10.311	6.893	-3.821	.549	-.022
11.1	.987	10.313	6.890	-3.820	.549	-.022
12.1	.987	10.313	6.890	-3.820	.549	-.022

Derivatives are calculated numerically.

a. Major iteration number is displayed to the left of the decimal, and minor iteration number is to the right of the decimal.

b. Run stopped after 12 iterations. Optimal solution is found.

表 10-65 中显示了渐近法和 150 次 Bootstrap 抽样计算出的各参数的估计值、标准误、95% 的置信区间、95% 的调整范围(它是将 150 个抽样估计值中最极端的 5% 剔除,即 P2.5~P97.5 的范围)。

表 10-65 参数估计

Parameter Estimates						
Parameter	Estimate	Std. Error	95% Confidence Interval		95% Trimmed Range	
			Lower Bound	Upper Bound	Lower Bound	Upper Bound
Asymptotic	a	10.313	7.951	-14.992	35.618	
	b	6.890	7.347	-16.490	30.271	
	c	-3.820	2.307	-11.163	3.524	
	d	.549	.297	-.398	1.496	
	e	-.022	.013	-.065	.021	
Bootstrap ^a	a	10.313	35.863	-60.554	81.180	-114.400
	b	6.890	27.314	-47.082	60.863	-19.967
	c	-3.820	7.703	-19.040	11.400	-25.115
	d	.549	.955	-1.339	2.436	-1.010
	e	-.022	.044	-.109	.065	-.110

a. Based on 150 samples.

由 Bootstrap 抽样法得到的计算结果可以被认为是确切结果,它可以用来和渐近的分析结果进行比较,从 5 个参数的估计值上来看,两种方法下得出的结果是完全一样的,但估计的标准误上,由 Bootstrap 抽样法得到的估计标准误要相应地大于渐近法得到的计算结果。应以 Bootstrap 抽样法得到的结果作为最终结果。

表 10-66 为渐近法和 Bootstrap 抽样法下计算得到的参数的相关系数矩阵。两者基本接近,但 Bootstrap 抽样法得到的结果要相应小于渐近法的计算结果。

表 10-67 给出了类似线性回归中的方差分析表, 所不同的是本表没有 F 值及其检验结果。因为这里是非线性回归, 它们只有参考意义, 但由这些结果, 我们也可以很容易得到参考的 F 检验的结果。

如 F 值为 $191.791/0.329=582.9514$, $p=0.0001$ 。

决定系数 R^2 为 0.995, 可以说拟合度达到了非常令人满意的程度。

表 10-66 相关系数矩阵

Correlations of Parameter Estimates					
	a	b	c	d	e
Asymptotic					
a	1.000	-.992	.974	-.951	.927
b	-.992	1.000	-.994	.981	-.963
c	.974	-.994	1.000	-.996	.986
d	-.951	.981	-.996	1.000	-.997
e	.927	-.963	.986	-.997	1.000
Bootstrap					
a	1.000	-.990	.960	-.914	.850
b	-.990	1.000	-.990	.960	-.911
c	.960	-.990	1.000	-.990	.958
d	-.914	.960	-.990	1.000	-.988
e	.850	-.911	.958	-.988	1.000

表 10-67 方差分析

ANOVA ^a			
Source	Sum of Squares	df	Mean Squares
Regression	958.953	5	191.791
Residual	.987	3	.329
Uncorrected Total	959.940	8	
Corrected Total	203.335	7	

Dependent variable: 铵态 N

a. R squared = $1 - (\text{Residual Sum of Squares}) / (\text{Corrected Sum of Squares}) = .995$.

由此, 我们可以得到所要求的四次多项式的回归方程为

$$\hat{y} = 10.313 + 6.890x - 3.820x^2 + 0.549x^3 - 0.022x^4$$

式中, \hat{y} 为铵态 N 含量的预测值, x 为 pH 含量。

10.5.4 多元线性回归分析

在实际问题中, 影响因变量的因素往往不止一个, 而是多个。在这种情形下, 为同一元线性回归分析有所区别, 而把研究因变量 y 与多个自变量 x_1, x_2, \dots, x_p 之间关系的问题称为多元回归分析。多元回归分析分为多元线性回归分析和多元非线性回归分析。研究因变量 y 与多个自变量 x_1, x_2, \dots, x_p 之间线性关系的问题称为多元线性回归分析。它同一元线性回归分析的基本原理是完全一样的, 只是计算更加复杂, 往往需要借助于计算机才能完成大量的运算工作。同一元非线性回归分析可以转化为一元线性回归分析一样, 由于多元非线性回归的问题大部分可以转化为多元线性回归的问题, 因此, 这里只介绍多元线性回归分析。

10.5.4.1 数学模型

1. 多元线性回归的数学模型

如果变量 y 与另外 p 个自变量 x_1, x_2, \dots, x_p 之间的内在联系是线性的, 它的第 α 次试验的数据是 $(y_\alpha; x_{\alpha 1}, x_{\alpha 2}, \dots, x_{\alpha p})$, $\alpha = 1, 2, \dots, n$, 则这一组数据可以假设有如下的结构式:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n \end{cases} \quad (10.9)$$

其中 $\beta_0, \beta_1, \dots, \beta_p$ 是 $p+1$ 个待估参数, x_1, x_2, \dots, x_p 是 p 个可以精确测量或可控制的一般变量, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 是 n 个相互独立且服从正态分布 $N(0, \sigma^2)$ 的随机变量。这就是多元线性回归的数学模型。称

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

为 Y 关于 x_1, x_2, \dots, x_p 的 p 元线性回归方程, 称 β_0 为回归常数, β_1, \dots, β_p 为偏回归系数。

令

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

则多元线性回归的数学模型用矩阵形式可以表示为

$$Y = X\beta + \varepsilon \quad (10.10)$$

其中 ε 是 N 维随机向量, 它的分量是相互独立的。

2. 参数 β 的 LS (最小二乘) 估计

设 b_0, b_1, \dots, b_p 分别是参数 $\beta_0, \beta_1, \dots, \beta_p$ 的 LS 估计, 则回归方程为

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \quad (10.11)$$

由最小二乘法可知, b_0, b_1, \dots, b_p 应使得全部观察值 y_α 与回归值 \hat{y}_α 的偏差平方和 Q 达到最小, 即

$$Q = \sum_{\alpha=1}^n (y_\alpha - \hat{y}_\alpha)^2 = \sum_{\alpha=1}^n (y_\alpha - b_0 - b_1 x_{\alpha 1} - b_2 x_{\alpha 2} - \cdots - b_p x_{\alpha p})^2 = \min$$

对于给定的数据, Q 是 b_0, b_1, \dots, b_p 的非负二次项, 所以最小值一定存在。根据微积分学中的极值原理, b_0, b_1, \dots, b_p 应是下列方程组的解

$$\begin{cases} \frac{\partial Q}{\partial b_0} = -2 \sum_{\alpha=1}^n (y_\alpha - \hat{y}_\alpha) = -2 \sum_{\alpha=1}^n (y_\alpha - b_0 - b_1 x_{\alpha 1} - b_2 x_{\alpha 2} - \cdots - b_p x_{\alpha p}) = 0 \\ \frac{\partial Q}{\partial b_j} = -2 \sum_{\alpha=1}^n (y_\alpha - \hat{y}_\alpha) x_{\alpha j} = -2 \sum_{\alpha=1}^n (y_\alpha - b_0 - b_1 x_{\alpha 1} - b_2 x_{\alpha 2} - \cdots - b_p x_{\alpha p}) x_{\alpha j} = 0 \end{cases} \quad (10.12)$$

其中, $j=1, 2, \dots, p$ 。

称方程组 (10.12) 为正规方程组, 它进一步可化为

$$\begin{cases} nb_0 + \left(\sum_{\alpha=1}^n x_{\alpha 1}\right)b_1 + \left(\sum_{\alpha=1}^n x_{\alpha 2}\right)b_2 + \cdots + \left(\sum_{\alpha=1}^n x_{\alpha p}\right)b_p = \sum_{\alpha=1}^n y_{\alpha} \\ \left(\sum_{\alpha=1}^n x_{\alpha 1}\right)b_0 + \left(\sum_{\alpha=1}^n x_{\alpha 1}^2\right)b_1 + \left(\sum_{\alpha=1}^n x_{\alpha 1}x_{\alpha 2}\right)b_2 + \cdots + \left(\sum_{\alpha=1}^n x_{\alpha 1}x_{\alpha p}\right)b_p = \sum_{\alpha=1}^n x_{\alpha 1}y_{\alpha} \\ \left(\sum_{\alpha=1}^n x_{\alpha 2}\right)b_0 + \left(\sum_{\alpha=1}^n x_{\alpha 1}x_{\alpha 2}\right)b_1 + \left(\sum_{\alpha=1}^n x_{\alpha 2}^2\right)b_2 + \cdots + \left(\sum_{\alpha=1}^n x_{\alpha 2}x_{\alpha p}\right)b_p = \sum_{\alpha=1}^n x_{\alpha 2}y_{\alpha} \\ \vdots \\ \left(\sum_{\alpha=1}^n x_{\alpha p}\right)b_0 + \left(\sum_{\alpha=1}^n x_{\alpha p}x_{\alpha 1}\right)b_1 + \left(\sum_{\alpha=1}^n x_{\alpha p}x_{\alpha 2}\right)b_2 + \cdots + \left(\sum_{\alpha=1}^n x_{\alpha p}^2\right)b_p = \sum_{\alpha=1}^n x_{\alpha p}y_{\alpha} \end{cases}$$

很明显, 正规方程的系数矩阵是对称矩阵。如果用 \mathbf{A} 来表示它, 则

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} n & \sum_{\alpha=1}^n x_{\alpha 1} & \sum_{\alpha=1}^n x_{\alpha 2} & \cdots & \sum_{\alpha=1}^n x_{\alpha p} \\ \sum_{\alpha=1}^n x_{\alpha 1} & \sum_{\alpha=1}^n x_{\alpha 1}^2 & \sum_{\alpha=1}^n x_{\alpha 1}x_{\alpha 2} & \cdots & \sum_{\alpha=1}^n x_{\alpha 1}x_{\alpha p} \\ \sum_{\alpha=1}^n x_{\alpha 2} & \sum_{\alpha=1}^n x_{\alpha 1}x_{\alpha 2} & \sum_{\alpha=1}^n x_{\alpha 2}^2 & \cdots & \sum_{\alpha=1}^n x_{\alpha 2}x_{\alpha p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum_{\alpha=1}^n x_{\alpha p} & \sum_{\alpha=1}^n x_{\alpha p}x_{\alpha 1} & \sum_{\alpha=1}^n x_{\alpha p}x_{\alpha 2} & \cdots & \sum_{\alpha=1}^n x_{\alpha p}^2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_{1p} & x_{2p} & x_{3p} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{np} & \cdots & x_{np} \end{pmatrix} \\ &= \mathbf{X}'\mathbf{X}。 \end{aligned}$$

同理, 正规方程组 (10.12) 右端的常数矩阵 \mathbf{B} 也可用矩阵 \mathbf{X} 和 \mathbf{Y} 表示

$$B = \begin{pmatrix} \sum_{\alpha=1}^n y_{\alpha} \\ \sum_{\alpha=1}^n x_{\alpha 1} y_{\alpha} \\ \sum_{\alpha=1}^n x_{\alpha 2} y_{\alpha} \\ \vdots \\ \sum_{\alpha=1}^n x_{\alpha p} y_{\alpha} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = X'Y$$

因此, 正规方程组 (10.12) 的矩阵形式是

$$(X'X)b = X'Y \quad (10.13)$$

或 $Ab = B$ 。

其中, $b' = (b_0, b_1, b_2, \dots, b_p)$, 是正规方程组 (10.13) 中的未知数。在系数矩阵 A 满秩的条件下, A 的逆矩阵 A^{-1} 存在, 因此,

$$b = A^{-1}B = (X'X)^{-1}X'Y \quad (10.14)$$

它就是模型 (10.10) 中参数的 LS 估计, 可以证明, 它是参数 β 的无偏估计, 有时也将它们统称为回归方程 (10.11) 的偏回归系数。用最小二乘估计法求出的诸偏回归系数相互之间存在相关性。

3. 回归方程的显著性检验

由于在求线性回归方程前, 我们一般不能断定随机变量 y 与一般变量 x_1, x_2, \dots, x_p 之间是否确有线性关系, 而是基于假设是呈线性关系的前提下做出的, 因此, 必须对做出的回归方程进行统计检验, 以便给出肯定或否定的结论。

对多元线性回归方程进行显著性检验的方法跟一元线性回归方程的显著性检验方法是一样的, 通常采用方差分析法。

设

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

是所求的回归方程, \hat{y}_{α} 是第 α 个试验点 $(x_{\alpha 1}, x_{\alpha 2}, \dots, x_{\alpha p})$ 上的回归值, 显然,

$$\hat{y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} = Xb = XA^{-1}X'Y$$

其总的偏差平方和

$$L_{\text{总}} = \sum_{\alpha=1}^n (y_{\alpha} - \bar{y})^2 = \sum_{\alpha=1}^n y_{\alpha}^2 - \frac{1}{n} \left(\sum_{\alpha=1}^n y_{\alpha} \right)^2$$

其自由度为: $f_{\text{总}} = n - 1$ 。

同一元线性回归一样, 可以证明

$$L_{\text{总}} = L_{\text{回}} + L_{\text{剩}}$$

其中

$$L_{\text{回}} = \sum_{\alpha=1}^n (\hat{y}_{\alpha} - \bar{y})^2, f_{\text{回}} = p$$

它是由于引入变量 x_1, x_2, \dots, x_p 以后引起的, 反映自变量对因变量的贡献大小。

$$L_{\text{剩}} = \sum_{\alpha=1}^n (y_{\alpha} - \hat{y})^2, f_{\text{剩}} = n - p - 1$$

它是由于试验误差和其他因素而引起的。

变量 y 与变量 x_1, x_2, \dots, x_p 之间是否存在线性关系, 实际上就是要检验原假设

$$H_0: \beta_1 = 0, \beta_2 = 0, \dots, \beta_p = 0$$

是否成立。也即检验

$$F = \frac{L_{\text{回}}/p}{L_{\text{剩}}/(n-p-1)} \sim F(p, n-p-1)$$

是否有显著性意义

对于事先给定的显著性水平 α , 当原假设成立的概率 $P(H_0) < \alpha$ 时, 拒绝原假设, 说明 y 与变量 x_1, x_2, \dots, x_p 之间存在线性关系, 否则, 认为 y 与变量 x_1, x_2, \dots, x_p 之间不存在线性关系。

4. 回归系数的显著性检验

回归方程显著, 说明 y 与变量 x_1, x_2, \dots, x_p 中至少有一个变量之间存在线性关系, 所以, 它并不意味着每个自变量 x_1, x_2, \dots, x_p 对变量 y 的影响都是重要的。因此, 为使所得到的回归方程更加实用, 能在实际中很方便地用来对 y 进行预测和控制, 在不影响预测精度的前提下, 我们总希望回归方程中的自变量越少越好, 因而需要把那些次要的、可有可无的自变量从回归方程中剔除。

显然, 如果某个自变量 x_j 对 y 的作用不显著, 则在多元回归方程中, 其前面的系数 β_j 就可以取值为零。因此, 检验因子 x_j 是否显著等价于检验假设

$$H_0: \beta_j = 0。$$

由于

$$F = \frac{(b_j - \beta_j)^2 / c_{jj}}{L_{\text{剩}} / (n - p - 1)} \sim F(1, n - p - 1)$$

或

$$t = \frac{b_j}{\sqrt{c_{jj} L_{\text{剩}} / (n - p - 1)}} = \frac{b_j}{\hat{\sigma} \sqrt{c_{jj}}} \sim t(n - p - 1)$$

式中, c_{jj} 为 A^{-1} 中对角线上第 j 个元素。

所以, 可用 F 或 t 统计量来检验回归系数是否显著。对不显著的自变量要从回归方程中剔除, 但每次剔除一个。

应该指出的是: 从回归方程中剔除一个变量, 这并不是把它这一项简单地从回归方程中拿去就可以了, 而是要重新从 $p-1$ 个变量着手, 估计偏回归系数, 写出新的回归方程。再重复上述检验过程, 直到余下的回归系数都显著为止。

设 $L_{\text{回}}$ 是 p 个变量 x_1, x_2, \dots, x_p 所引起的回归平方和, 它是所有自变量对 y 波动的总贡献, $L'_{\text{回}}$ 是 $p-1$ 个变量 $x_1, x_2, x_{i-1}, \dots, x_p$ 所引起的回归平方和 (即除去 x_i), 那么它们的差

$$Q_i = L_{\text{回}} - L'_{\text{回}} = \frac{b_i^2}{c_{ii}}$$

就是去掉变量 x_i 后, 回归平方和减少的量。将其称为**偏回归平方和**。偏回归平方和越大, 表示该变量在回归中起的作用越大。在回归方程显著的前提下, 至少有一个变量 x_i 与 y 之间存在线性关系, 因此, 偏回归平方和最大的变量, 一定是显著的; 而偏回归平方和小的变量, 却并不一定不显著, 但可以肯定, 偏回归平方和最小的变量, 必然是所有变量中对 y 作用最小的一个。当该变量检验结果不显著时, 就可以将该变量剔除。

5. 建立多元线性回归方程的方法有

[方法 1] 将所需要的自变量强制性地选入到回归方程中。然后, 从所有可能的变量组合的回归方程中挑选最优者。

[方法 2] 先将全部变量选入到回归方程中, 然后从包含所有变量的回归方程中逐次剔除不显著的自变量。

[方法 3] 先选一个与因变量线性相关系数最大的变量进入到回归方程中, 再把同因变量相关系数较大的变量逐个引入回归方程。

[方法 4] 结合方法 3 和方法 2 进行边进边出的逐步回归法。

6. 关于一个好的回归方程选择的标准

一个好的回归方程, 首先应满足在回归方程中要包含所有对 y 影响显著的变量而不包含对 y 影响不显著的变量。

此外, 还应满足以下一些条件:

(1) 要有较大的决定系数 R^2 尤其是校正 R^2 。

在一元线性回归分析中, 已经提到过, 决定系数 $R^2 = \frac{L_{\text{回}}}{L_{\text{总}}}$ 可以用来反映用自变量大约

可以解释因变量变化的百分比, 考虑到由于 R^2 有当自变量数目增加而增大的缺点, 因此, 在多元线性回归中, 通常用校正 R^2

$$\text{Adjusted } R^2 = 1 - \frac{\sum_{i=1}^n (y - \hat{y})^2 / (n - k - 1)}{\sum_{i=1}^n (y - \bar{y})^2 / (n - 1)}$$

(式中, k 为自变量的个数, n 为样本容量)

来说明自变量大约可以解释因变量变化的百分比。

所以, 校正 R^2 的大小被通常用来评判多元线性回归方程的效果。因此, 一个好的回归方程应是方程中包含的自变量都是对因变量有显著影响的, 同时还具有较高的校正 R^2 值的回归方程。

(2) 要有较小的剩余标准差 S_e 或均方误差 (MSE) (S_e^2)。

剩余标准差是残差的标准差, 可以用来反映回归方程的估计精度, 其平方被称为残差的方差, 又称为均方误差, 其值越小越好, 它会随回归方程中自变量的增加而减小, 但当在回归方程中增加一些无统计意义的自变量后, 剩余标准差的值会不减反增。这一性质同校正 R^2 类似。

(3) 较小的 AIC 值

AIC 又称赤池信息准则, 它是日本学者赤池 (Akaike) 于 1973 年提出的, 广泛用于时间序列分析中自回归阶数的确定、多元回归、广义线性回归中自变量的筛选等。

当模型或方程是用 LS 估计时, AIC 定义为

$$AIC = n \ln((n - p) / n \times s_e^2) + 2p$$

当模型或方程是用极大似然法估计时, AIC 定义为

$$AIC = -2 \ln(L) + 2p$$

式中, p 是模型中参数的个数, L 是模型的极大似然函数, n 为样本含量。AIC 由两部分组成, 前面一部分反映回归方程的拟合精度, 其值越小越好, 第二部分反映回归中变量的多少, 即模型的复杂程度, 因此, 也是越小越好, 故两部分相加而得的 AIC 也是越小越好。

(4) C_p 应接近于 P

C_p 统计量由 C.L.Mallows 于 1964 年提出, 其定义如下

$$C_p = \frac{SSE_p}{MSE_m} - (n - 2p) = \frac{(n - p - 1)MSE_p}{MSE_m} - (n - 2p)$$

式中, MSE_p 是只有 p 个变量时的模型的残差平方和, MSE_m 是全部变量都在方程中时的均方误差。

“最优”方程是所有可能的回归方程中, C_p 最接近于 P 的那个回归方程。

7. 多重共线性的诊断及解决办法

多重共线性是指自变量间存在近似的线性关系, 也就是一个自变量能近似地用其他自变量的线性函数来描述。在多元线性回归分析中, 由于要考虑的自变量是多个, 因此, 自变量之间要达到完全独立是很难做到的, 低度的相关对回归结果的影响不会太大, 但当变量间共线性趋势非常严重时, 会对模型的拟合带来严重的影响。

(1) 多重共线性的诊断

① 检查自变量间的线性相关系数矩阵, 观察变量间的线性相关系数是否有很高的值, 一般线性相关系数的值超过 0.8 时, 就可能有共线性问题的存在, 超过 0.9 时共线性问题会变得比较严重。这只是初步的诊断, 并不全面。

② 检查容忍度 (Tolerance) 的值

容忍度统计量由 Norusis 等提出, 它是以每个自变量作为应变量对其他自变量进行回归分析时得到的残差比例, 其计算公式为

$$Tol_i = 1 - R_i^2$$

其值越小, 说明自变量 x_i 与其他自变量之间的共线性越强。

使用容忍度作为共线性量度的指标时, 要求变量的观测值要服从或近似服从正态分布。

③ 方差膨胀因子 (VIF)

它由 Marguardt 于 1960 年提出, 其计算公式为

$$VIF_i = \frac{1}{Tol_i}$$

即容忍度的倒数。 VIF 值越大, 说明共线性问题越严重。

④ 条件指数 (Condition Index)

它由 Stewart 等提出, 其计算公式为

$$Condition\ Index = \sqrt{\lambda_{\max} / \lambda_i}$$

式中, λ 为特征值。计算得到的条件指数的值越大, 说明自变量之间有共线性的可能性越大。有些学者认为, 当某些维度的该指标值大于 30 时, 则可能存在共线性, 但理论上并没有得到证明。

⑤ 方差比例

如果同一序号的特征值对应的变量的方差比例越大,说明存在共线性的可能性越大。

⑥ 特征值 (Eigenvalue): 对自变量进行主成分分析,如果相当多维度的特征值约等于 0,则可能存在比较严重的共线性。

(2) 解决办法

用存在共线性的变量直接进行多元线性回归分析肯定是不行的,此时,采用的解决办法有:

① 增大样本含量,可以部分解决共线性问题。

② 利用对自变量进行聚类分析选择各类典型指标的方法来减少自变量之间的相关程度。

③ 对自变量进行主成分分析,用提取的因子代替原变量进行回归分析,可极大地消除共线性问题。

④ 进行岭回归、通径分析等也可有效地解决多重共线性问题。

⑤ 从专业的角度,剔除高度相关的自变量中的一些次要的变量,使其不选入回归方程。

8. 进行多元线性回归分析时另外需要注意的问题

(1) 要对用来进行回归分析的数据资料进行审核,对特大或特小的异常值,有条件的要通过重测来加以确认,或从专业角度来对其进行判别,确认是异常值的要从原始数据中剔除。

(2) 从多元线性回归的原理上而言,样本含量 n 至少应大于 $p+1$,根据人们的经验,样本含量至少为自变量数 p 的 5 倍,也有学者提出应至少是自变量数 p 的 20 倍。

(3) 因变量要服从或近似服从正态分布。对于不服从正态分布的,要进行预处理,进行必要的变量变换,以确保其正态性和方差齐性。对于用转换后服从正态分布的因变量所做的回归方程,在求得它的预测值后还要还原成最初因变量的形式。

9. 利用回归方程进行预测和控制

如果得到的多元线性回归方程经检验回归显著时,就可以用所得的经验回归方程作预测。

同一元线性回归作预测时一样,当取得 x_1, x_2, \dots, x_p 的一组观测值 $x_{01}, x_{02}, \dots, x_{0p}$ 时,则 y_0 的点估计值为

$$\hat{y}_0 = b_0 + b_1 x_{01} + \dots + b_p x_{0p}$$

同样,理论上可以证明

$$T = \frac{y_0 - \hat{y}_0}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \sum_{i=1}^p \sum_{j=1}^p c_{ij} (x_{0i} - \bar{x}_i)(x_{0j} - \bar{x}_j)}} \sim t(n-p-1)$$

其中, $\hat{\sigma} = s_e = \sqrt{\frac{L_{\text{剩}}}{n-p-1}}$, c_{ij} 为 A^{-1} 的元素, n 为样本含量, p 为回归方程中的自变量数。

对于给定的置信度 $100(1-\alpha)\%$, 可得 y_0 的 $100(1-\alpha)\%$ 的预测区间为

$$[\hat{y}_0 - t_{1-\alpha/2}(n-p-1)\delta, \hat{y}_0 + t_{1-\alpha/2}(n-p-1)\delta]$$

$$\text{其中, } \delta = \sqrt{\frac{L_{\text{剩}}}{n-p-1}} \cdot \sqrt{1 + \frac{1}{n} + \sum_{i=1}^p \sum_{j=1}^p c_{ij}(x_{0i} - \bar{x}_i)(x_{0j} - \bar{x}_j)}$$

当 n 较大且 x_{0j} 接近于 \bar{x}_j ($j=1, 2, \dots, p$) 时, 可用下式进行预测和控制

$$P\{\hat{y}_0 - 2\hat{\sigma} < y_0 < \hat{y}_0 + 2\hat{\sigma}\} = 95\%$$

$$P\{\hat{y}_0 - 3\hat{\sigma} < y_0 < \hat{y}_0 + 3\hat{\sigma}\} = 99\%。$$

10.5.4.2 实例分析

1. 自变量都为定量变量的线性回归分析

例 10.19 铝合金化学铣切工艺中, 为了便于生产操作, 需要对腐蚀速度进行控制, 为考察腐蚀液温度 ($^{\circ}\text{C}$) x_1 、碱浓度 (g/L) x_2 、腐蚀液含铝量 (g/L) x_3 对腐蚀速度 (mm/min) y 的影响, 共做了 44 次试验, 试验结果的数据资料已存放在 data10-15.sav 中, 试对其做多元线性回归分析, 并求 $x_{01} = 82^{\circ}\text{C}$, $x_{02} = 35 \text{ g/L}$, $x_{03} = 200 \text{ g/L}$ 时, 腐蚀速度 y_0 的点预测与 99% 的预测区间。

在 SPSS 中的解题步骤如下:

- ① 在数据编辑窗口中, 打开 data10-15.sav。
- ② 对因变量腐蚀速度做正态性检验

用第 2 章数据资料探索性分析中介绍的方法, 可得腐蚀速度变量的正态性检验结果, 见表 10-68。从表中可见, 腐蚀速度观察的显著性水平为大于等于 0.200, 大于 0.05, 故现有证据不足于拒绝腐蚀速度变量服从正态分布的原假设。

表 10-68 腐蚀速度变量的正态性检验

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
腐蚀速度	.079	44	.200 [*]	.977	44	.511

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

- ③ 用逐步回归法做多元线性回归分析

按 Analyze→Regression→Linear 顺序展开 Linear Regression 对话框, 见图 10-10。在左侧变量名源框中, 选择腐蚀速度变量, 将其移入到 Dependent 框中, 作为因变量, 选择温度、碱浓度、含铝量变量, 将其移入到 Independent(s)框中, 作为自变量。

在 Method 的下拉式选项中，选择 Stepwise 选项，要求做逐步回归分析。
单击 Statistics 按钮，弹出 Linear Regression: Statistics 对话框，见图 10-30。

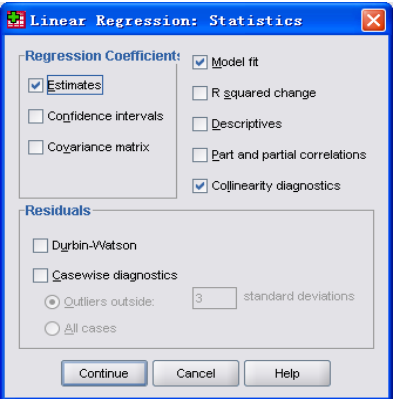


图 10-30 Linear Regression: Statistics 对话框

选择 Normal probability plot 选项，要求输出标准化残差的正态概率图。

其他保持系统默认选择。

单击 Continue 按钮，返回到 Linear Regression 对话框。

单击 OK 按钮运行，在输出窗口中得到七张计算结果表和两张残差图，分别参见表 10-69 至表 10-75 和图 10-31、图 10-32。

④ 结果与分析

表 10-69 列出了在 3 种模型中，自变量被引进或从回归方程中剔除的情况及采用的方法（逐步回归）。此表在实际应用中一般没多大用处，可以不用将它放在论文或科研成果中。

表 10-69 变量引进或剔除

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	温度	.	Stepwise (Criteria: Probabil- ity-of-F- to-enter ≤ .050, Probabil- ity-of-F- to- remove ≥ .100).
2	碱浓度	.	Stepwise (Criteria: Probabil- ity-of-F- to-enter ≤ .050, Probabil- ity-of-F- to- remove ≥ .100).
3	含铝量	.	Stepwise (Criteria: Probabil- ity-of-F- to-enter ≤ .050, Probabil- ity-of-F- to- remove ≥ .100).

a. Dependent Variable: 腐蚀速度

表 10-70 模型摘要

Model Summary ^d				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.895 ^a	.802	.797	.0031231
2	.953 ^b	.909	.904	.0021444
3	.976 ^c	.952	.949	.0015695

a. Predictors: (Constant), 温度

b. Predictors: (Constant), 温度, 碱浓度

c. Predictors: (Constant), 温度, 碱浓度, 含铝量

d. Dependent Variable: 腐蚀速度

表 10-70 为模型摘要, 从左向右各列依次为: 模型种类, 复相关系数 R 、决定系数 R^2 、校正 R^2 和估计的标准误。从上到下, 它共列出了三个模型下的各个统计量的值。三个模型中的因变量都为腐蚀速度。在第一个模型中, 自变量只有温度一个。在第二个模型中自变量为温度和碱浓度两个。第三个模型是包含全部三个自变量的情形。从校正 R^2 来看, 含有三个自变量的模型有最大的校正 R^2 值, 它为 0.952, 说明用三个自变量组成的线性组合大约可以解释 95.2% 的因变量的变化, 表明在因变量的变化中它们起主要作用。

表 10-71 方差分析表

ANOVA^d

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.002	1	.002	169.894	.000 ^a
	Residual	.000	42	.000		
	Total	.002	43			
2	Regression	.002	2	.001	204.219	.000 ^a
	Residual	.000	41	.000		
	Total	.002	43			
3	Regression	.002	3	.001	266.318	.000 ^a
	Residual	.000	40	.000		
	Total	.002	43			

a. Predictors: (Constant), 温度
b. Predictors: (Constant), 温度, 碱浓度
c. Predictors: (Constant), 温度, 碱浓度, 含铝量
d. Dependent Variable: 腐蚀速度

表 10-72 回归系数表

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-.074	.008		-9.084	.000		
	温度	.001	.000	.895	13.034	.000	1.000	1.000
2	(Constant)	-.082	.006		-14.387	.000		
	温度	.001	.000	1.022	20.206	.000	.870	1.150
	碱浓度	.000	.000	-.351	-6.934	.000	.870	1.150
3	(Constant)	-.094	.005		-20.351	.000		
	温度	.001	.000	1.022	27.607	.000	.870	1.150
	碱浓度	.000	.000	-.351	-9.474	.000	.870	1.150
	含铝量	6.000E-5	.000	.209	6.044	.000	1.000	1.000

a. Dependent Variable: 腐蚀速度

表 10-71 中的方差分析表的结构和内容同一元线性回归中的方差分析表实质是一样的, 唯一不同的是本处是三个模型的叠加表。

从表 10-71 可见, 三个模型均具有统计上的极显著意义(P 都接近于 0.000 小于 0.01), 说明 3 个自变量各自及其组合与因变量之间都存在线性回归关系, 结合表 10-46 中得出的结果, 在三个模型中, 相对而言, 模型 3 为最佳。其残差均方 MSE 为 2.4635×10^{-6} 。剩余标准差 $s_e = \sqrt{2.4635 \times 10^{-6}} = 0.00157$, 它实际上就是表 10-65 模型 3 中的估计的标准误。

表 10-72 列出了偏回归系数及其分析结果。从左向右依次为: 模型结构、原始数据下的回归系数、回归系数的标准误、标准化回归系数、对回归系数检验的 t 值、观测值的显著性水平、容忍度和方差膨胀因子。

由于上面的讨论已将重点关注在模型 3 上, 因此, 对此表分析的重点也放在模型 3 部分。从该表可知, 所要求的多元线性回归方程为

$$\hat{y} = -0.09432 + 0.001479x_1 - 2.16035 \times 10^{-4}x_2 + 6.000 \times 10^{-5}x_3$$

从 Sig. 列可以看出, 方程中的常量和各个自变量对应的回归系数均有统计上的显著性意义 (P 都接近于 0.000 小于 0.01)。

从容忍度指标来看, 由于三个自变量的容忍度都在 0.870 以上接近于 1, 因此, 在本例中不用担心多重共线性问题。方差膨胀因子是容忍度的倒数, 因此由它得到的结论同容忍度是一样的。

在你的分析中, 这一步的分析要先进行, 因为, 当多重共线性存在时, 上述的方程无多大实用意义, 还要等剔除了部分共线性严重的次要变量后, 才可能得到类似上述有意义的方程。

表 10-73 显示在模型中尚未包含变量检验的一些基本情况。从左向右依次为: 模型中尚未包括的变量、变量对应的标准回归系数、 t 统计量值、观测值的显著性水平、偏相关系数、容忍度、方差膨胀因子、最小容忍度。

从表 10-73 可见, 在模型 1 中, 尚未引入回归方程的碱浓度和含铝量变量都有统计学意义 ($P < 0.01$), 从容忍度和方差膨胀因子来看, 都不存在明显的共线性, 表示它们还都有可能在下一步中引入到模型中。在模型 2 中, 尚未引入回归方程的含铝量变量有统计学意义 ($P < 0.01$), 从容忍度和方差膨胀因子来看, 不存在明显的共线性, 表示它还有可能在下一步中引入到模型中。

表 10-74 给出的是对自变量进行主成分分析后的特征根和条件指数。虽然从条件指数上来看, 某些维度出现了条件指数大于 30 的情形, 当从方差比例来看, 三个变量方差比大的都不在同一维度上, 分布比较均匀, 大部分特征值都不约等于 0, 因而也没有足够的证据认为有明显的共线性的存在。

表 10-73 尚未引入回归方程的变量

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	碱浓度	-.351*	-6.934	.000	-.735	.870	1.150	.870
	含铝量	.209*	3.398	.002	.469	1.000	1.000	1.000
2	含铝量	.209*	6.044	.000	.691	1.000	1.000	.870

a. Predictors in the Model: (Constant), 温度

b. Predictors in the Model: (Constant), 温度, 碱浓度

c. Dependent Variable: 腐蚀速度

表 10-74 共线性诊断

Model	Dimensionality	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	温度	碱浓度	含铝量
1	1	1.998	1.000	.00	.00		
	2	.002	34.587	1.00	1.00		
2	1	2.925	1.000	.00	.00	.01	
	2	.074	6.295	.01	.01	.90	
	3	.002	43.735	.99	.99	.09	
3	1	3.903	1.000	.00	.00	.01	.00
	2	.086	6.756	.00	.00	.87	.02
	3	.010	20.158	.04	.07	.04	.91
	4	.001	52.034	.96	.93	.08	.07

a. Dependent Variable: 腐蚀速度

表 10-75 是残差统计结果表, 列出了预测值、标准预测值、预测值标准误等一系列指标的最小值、最大值、均值、标准差和样本量等统计结果, 对分析作用不大, 可忽略。

表 10-75 残差统计量

Residuals Statistics ^a					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.018266	.043797	.032011	.0067655	44
Std. Predicted Value	-2.032	1.742	.000	1.000	44
Standard Error of Predicted Value	.000	.001	.000	.000	44
Adjusted Predicted Value	.018216	.043756	.032012	.0067574	44
Residual	-3.519E-3	.0036411	-1.63E-18	.0015138	44
Std. Residual	-2.242	2.320	.000	.964	44
Stud. Residual	-2.342	2.437	.000	1.014	44
Deleted Residual	-3.838E-3	.0040194	-2.707E-7	.0016749	44
Stud. Deleted Residual	-2.489	2.608	-.001	1.041	44
Mahal. Distance	.100	6.857	2.932	1.944	44
Cook's Distance	.000	.154	.027	.038	44
Centered Leverage Value	.002	.159	.068	.045	44

a. Dependent Variable: 腐蚀速度

Normal P-P Plot of Regression Standardized Residual

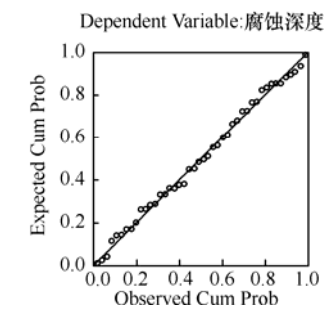


图 10-31 回归标准残差的正态概率图

Scatterplot

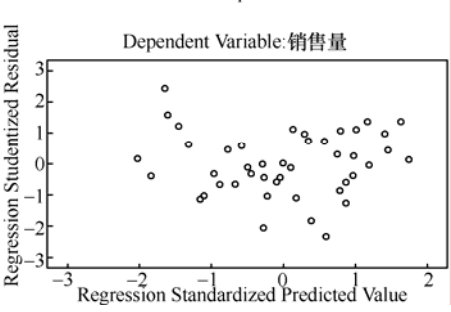


图 10-32 标准回归预测值与学生化残差的散点图

从图 10-31 可见，回归标准残差几乎都在 P-P 图的对角线线上，因而，可以认为残差是服从或近似服从正态分布的。

图 10-32 显示了残差的散点分布，由于这些点云呈水平带状，散开分布，没有形成一个明显的图案，表明数据满足误差是正态分布和残差的方差是常数的假定。

⑤ 预测

将 $x_{01} = 82\text{ }^{\circ}\text{C}$ ， $x_{02} = 35\text{ g/L}$ ， $x_{03} = 200\text{ g/L}$ 代入上面的预测方程中，可以得到 y_0 的点估计 $\hat{y} = 0.031397$ ，又 $s_e = 0.00157$ ，可以验证， $x_{01} = 82\text{ }^{\circ}\text{C}$ 和 $x_{03} = 200\text{ g/L}$ 非常接近于其变量的均值，而 $n = 44$ 也算较大，所以，可用

$$P\{\hat{y}_0 - 3\hat{\sigma}\langle y_0 \rangle \hat{y}_0 + 3\hat{\sigma}\} = 99\%$$

来作为 y_0 的 99% 的预测区间，代入得 y_0 的 99% 的预测区间为：[0.026687, 0.036107]。

2. 自变量中有定性变量的线性回归分析

当自变量中有定性变量（如性别、民族、职业、文化程度、季节等）时，一般首先将定性变量作定量处理，如引入数量“1”表示性别“男”，“0”表示性别“女”等，称对定性变量进行量化处理的变量为虚拟变量或哑元（dummy variable）。只取 0，1 两

个值的虚拟变量称 0—1 型虚拟变量。

虽然虚拟变量取某一个数值,但这个数值没有任何数量大小的意义,它仅仅是用来说明观察单位的性质和属性的。

对含有虚拟变量的多元线性回归模型的构造要比只有定量变量时麻烦得多。一般而言,对于一个具有 k 种特征或状态或称 k 个水平的定性变量回归模型,若回归模型不带常数项,则需引入 k 个 0—1 型虚拟变量 D ;若有常数项,只需引入 $k-1$ 个 0—1 型虚拟变量。当 $k=2$ 时,只需引入一个 0—1 型虚拟变量。模型的构造,不但同虚拟变量所分的水平数有关,还同虚拟变量的数量有关。当虚拟变量不止一个时,若它们之间还有交互作用,则在模型中还要考虑它们之间的交互作用,正如在多因素方差分析中提到的一样,每个交互作用项也需要引入一个 0—1 型虚拟变量。引进满足上述条件的 0—1 型虚拟变量的模型设定后,才可以用上面 1. 中的多元线性回归模型进行回归分析。

例如,要建立影响空调销售额的多元线性回归分析,除采用定量变量:当地居民的收入 x_1 、空调价格 x_2 作为自变量外,还要考虑定性变量:季节 D_1 (1-夏季, 0-冬季) 和地区 D_2 (1-沿海地区, 0-内地城市) 对空调销售额的影响,同时还认为这两个定性变量之间有交互作用,则此时要建立的一个空调销售额的回归模型为

$$y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_1 D_2 + \beta_4 x_1 + \beta_5 x_2 + \varepsilon$$

式中, β_0 为回归常数项, $\beta_1, \beta_2, \dots, \beta_5$ 为偏回归系数。 $D_1 D_2$ 为交互作用项,需生成一个新的 0—1 型虚拟变量。

由此可见,模型中的参数不再是固定不变的,而是随着虚拟变量数的增加和水平数的变化而发生变化。如果能事先确定回归的模型,则依然可以按 1. 中的做法进行多元线性回归分析。

• 定性变量为 0—1 型虚拟变量时多元线性回归分析

例 10.20 试用存放在 data10-16.sav 中的数据资料,建立保险公司的规模及其类型对采取某项革新措施的速度影响的多元线性回归方程。

这是一个自变量为定量(规模 x_1) 变量加定性(类型 D (0-股份公司, 1-有限责任公司)) 的 0—1 虚拟变量的情形,因此,需要建立的回归的模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D + \varepsilon$$

故,它可以在线性回归程序中进行处理。

表 10-76 速度变量的正态性检验

Tests of Normality						
速度	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
速度	.091	20	.200 [*]	.984	20	.974

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

在 SPSS 中的解题步骤如下:

1. 在数据编辑窗口中,打开 data10-16.sav。

2. 对因变量速度做正态性检验。

用第 2 章数据资料探索性分析中介绍的方法,对速度变量进行正态性检验,其结果

见表 10-76。从表中可见，速度观察的显著性水平为大于等于 0.200，大于 0.05，故现有证据不足于拒绝速度变量服从正态分布的原假设。

3. 用强制进入法 (Enter) 做多元线性回归分析。

按 Analyze→Regression→Linear 顺序展开 Linear Regression 对话框，见图 10-10。在左侧变量名源框中，选择速度变量，将其移入到 Dependent 框中，作为因变量，选择规模、类型变量，将其移入到 Independent(s)框中，作为自变量。

在 Method 的下拉式选项中，选择 Enter 选项，要求做所有自变量强制进入模型的回归分析。

按 Continue 按钮，返回 Linear Regression 对话框。

按 Plots 按钮，打开 Linear Regression: Plots 对话框，见图 10-13，在左侧变量框中选择 ZPRED 选项将其移入到 X 轴的框中，用标准预测值作为 X 轴变量，选择 SRESID 选项将其移入到 Y 轴的框中，用学生化残差作为 Y 轴变量。

在 Standardized Residual Plots 选项中，选择 Normal probability plot 选项，要求输出标准化残差的正态概率图。

其他保持系统默认选择。

按 Continue 按钮，返回到 Linear Regression 对话框。

按 OK 按钮运行，在输出窗口中得到六张计算结果表和两张残差图，分别参见表 10-77 至表 10-81 和图 10-33、图 10-34。

4. 结果与分析。

从表 10-77 可见，本例采用了强制性进入法 (Enter) 进行回归分析，进行模型的自变量有类型和规模。

从表 10-78 可见，自变量的线性组合与因变量间的复相关系数为 0.946， $R^2 = 0.895$ ，校正 $R^2 = 0.883$ ，表明自变量的线性组合可以解释因变量 88.3% 的变化，估计的标准误差为 3.221。

表 10-77 变量的引入或删除

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	类型, 规模 ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: 速度

表 10-78 模型摘要

Model Summary ^a				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.946 ^a	.895	.883	3.221

a. Predictors: (Constant), 类型, 规模

b. Dependent Variable: 速度

从表 10-79 可见，由规模和类型所做的自变量对因变量速度的线性回归，在统计学上有极显著性意义 ($P=0.000<0.01$)。

从表 10-80 可见，两个自变量的容忍度都为 0.975 接近于 1，表明自变量间基本不存在共线性的影响，而且，回归系数的检验结果也表明，常数项有统计学上的极显著性意

义 ($P=0.000<0.01$), 两个自变量所对应的回归系数也有统计学上的极显著性意义 ($P=0.000<0.01$), 表明所得到的回归方程是“较优”的。

表 10-79 方差分析表

ANOVA ^a					
Model		Sum of Squares	df	Mean Square	F
1	Regression	1504.413	2	752.207	72.497
	Residual	176.387	17	10.376	
	Total	1680.800	19		

a. Predictors: (Constant), 类型, 规模

b. Dependent Variable: 速度

表 10-80 回归系数表

Coefficients ^a							
Model		Unstandardized Coefficients		Standardized Coefficients		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance VIF ^b
1	(Constant)	33.874	1.814		18.675	.000	
	规模	-.102	.009	-.911	-11.443	.000	.975 1.026
	类型	8.055	1.459	.439	5.521	.000	.975 1.026

a. Dependent Variable: 速度

表 10-81 的含义同表 10-75, 对讨论的意义不大, 本例不再解释。

表 10-81 残差统计量

Residuals Statistics ^a					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	4.37	35.01	19.40	8.898	20
Std. Predicted Value	-1.689	1.754	.000	1.000	20
Standard Error of Predicted Value	1.020	1.593	1.235	.181	20
Adjusted Predicted Value	4.47	34.15	19.41	8.862	20
Residual	-5.692	6.341	.000	3.047	20
Std. Residual	-1.767	1.968	.000	.946	20
Stud. Residual	-1.962	2.118	-.002	1.022	20
Deleted Residual	-7.020	7.343	-.012	3.558	20
Stud. Deleted Residual	-2.165	2.395	.002	1.087	20
Mahal. Distance	.956	3.699	1.900	.852	20
Cook's Distance	.001	.300	.056	.081	20
Centered Leverage Value	.050	.195	.100	.045	20

a. Dependent Variable: 速度

图 10-33 残差的正态概率图表明, 残差都在对角线上、下波动, 所以它基本是服从正态分布的。而残差的散点图 (见图 10-34) 的点云呈水平带状, 散开分布, 没有呈明显的图形, 表明数据满足误差是正态分布和残差的方差是常数的假定的。

Normal P-P Plot of Regression Standardized Residual

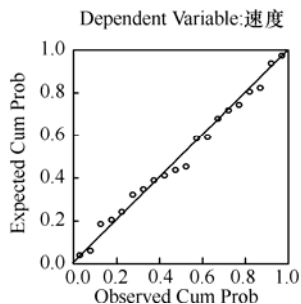


图 10-33 回归标准残差的正态概率图

Scatterplot

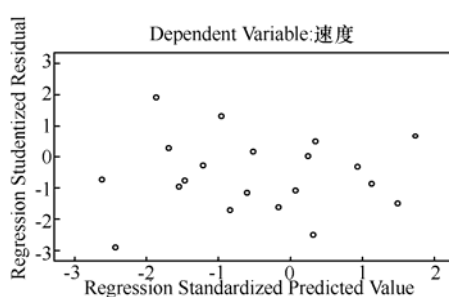


图 10-34 标准回归预测值与学生化残差的散点图

由此, 可以得到显著的经验回归方程为

$$\hat{y} = 33.874 - 0.102x_1 + 8.055D$$

实际上, 上面的回归方程可以表述为两个回归方程

对于股份制保险公司采取某项革新措施的速度可用

$$\hat{y} = 41.929 - 0.102x_1$$

来预测。

而对于有限责任公司采取某项革新措施的速度可用

$$\hat{y} = 33.874 - 0.102x_1$$

来预测。

- 对于定性变量分类数大于 2 的虚拟变量的线性回归分析

在定性变量的分类数 $k > 2$ 时，虽然也可以按上述的做法来进行回归分析，但需要将多分类变量变成多个 0—1 型虚拟变量，因此，这样做是非常麻烦的。此时，应用 SPSS 的单因变量多因素方差分析程序进行处理，就不需要先将一个多分类的定性变量，先转换成多个 0—1 型的虚拟变量了。

例 10.21 为研究在相同密度下每穗粒数、百粒重量、株高类别与玉米子粒产量的关系，测试了 12 个北方春玉米杂交种的数据，见 data10-17.sav，试用每穗粒数、百粒重量和株高类别作为自变量建立对玉米子粒产量的最优多元线性回归方程。

【题析】 由于株高类别按玉米株高度共分三类，1—大于等于 290 厘米，2—大于等于 282 厘米且小于 290 厘米，3—小于 282 厘米（见数据文件）。所以，它是分类数大于 2 的虚拟变量，不能直接用例 10.20 中的做法进行多元线性回归分析。

本例的具体解题步骤如下：

1. 产量变量的正态性检验。

首先，在数据编辑窗口中，打开数据文件 data10-17.sav。利用第 2 章 2.4 探索分析例 2.48 中介绍的数据资料的正态性和方差齐性检验方法，对产量变量进行正态性检验，可得检验结果，见表 10-82。

表 10-82 玉米产量变量的正态性检验

Tests of Normality						
产量	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
	.149	12	.200 [*]	.956	12	.728

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

从表 10-82 可见，因观测值的显著性水平为大于 0.200，故不拒绝产量变量服从正态分布的假设。

2. 进行多元线性回归分析。

按 Analyze→General Linear Model→Univariate 顺序，打开 Univariate 主对话框，见图 8-1。在左侧源变量框中，选择产量变量，单击向右箭头，将其送入 Dependent Variable 框；用同样方法分别将株高变量送入 Fixed Factor (s) 框中，将每穗总粒数、百粒重变量送入 Covariate (s) 框中。

单击 Model 按钮，在弹出的 Univariate:Model 对话框中，在 Specify Model 项中选择

Custom 选项, 对模型进行自定义, 在 Build Term(s) 的 Type 下拉式选项中, 选择 Main effects 选项, 要求模型中只包含主效应。关闭 Include intercept in model 选项, 其他采用系统默认选项, 按 Continue 按钮, 返回 Univariate 主对话框。

单击 Options 按钮, 在弹出的 Univariate: Options 对话框中, 在 Display 项中选择 Parameter estimates 选项, 要求输出参数估计。

按 OK 按钮运行, 则在输出窗中, 得到三张表, 见表 10-83、表 10-84 和表 10-85。

3. 结果与讨论。

表 10-83 显示的是株高虚拟变量的基本情况, 它分三个类别, 各类的定义及出现的观测的次数。该表在论文中一般不用。

表 10-84 给出了对模型及模型中变量的检验结果。方差检验的结果表明, 模型是显著的, 三个变量中除株高虚拟变量在 0.10 水平显著外 ($P=0.094$), 另外两个变量均在 0.01 水平上显著 ($P=0.002$ 和 0.000)。

表 10-83 对株高虚拟变量的描述

Between-Subjects Factors		
	Value Label	N
株高	1 大于等于 290 厘米	3
	2 大于等于 282 且小于 290 厘米	5
	3 小于 282 厘米	4

表 10-84 模型检验

Tests of Between-Subjects Effects					
Dependent Variable: 产量					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	8.778E8 ^a	5	1.756E8	1.514E3	.000
株高	1102915.332	3	367638.444	3.171	.094
每穗总粒数	2840949.855	1	2840949.855	24.504	.002
百粒重	6078867.856	1	6078867.856	52.432	.000
Error	811570.321	7	115938.617		
Total	8.786E8	12			

a. R Squared = .999 (Adjusted R Squared = .998)

表 10-85 给出了模型中的参数及其检验结果。从表中可见 5 个参数均在 0.05 水平上显著。

表 10-85 参数估计

Parameter Estimates						
Dependent Variable: 产量						
Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
[株高=1]	-7.271E3	2420.640	-3.004	.020	-12995.343	-1547.532
[株高=2]	-7.019E3	2410.583	-2.912	.023	-12719.508	-1319.260
[株高=3]	-6.831E3	2335.114	-2.925	.022	-12352.453	-1309.120
每穗总粒数	14.516	2.932	4.950	.002	7.582	21.450
百粒重	240.464	33.209	7.241	.000	161.938	318.990

因此, 所要求的最佳回归方程为:

适用于株高大于等于 290 厘米时, 预测玉米产量的经验回归方程

$$\hat{y} = 14.516x_1 + 240.464x_2 - 7.271 \times 10^3$$

适用于株高大于等于 282 厘米且小于 290 厘米时, 预测玉米产量的经验回归方程

$$\hat{y} = 14.516x_1 + 240.464x_2 - 7.019 \times 10^3$$

以及适用于株高小于 282 厘米时, 预测玉米产量的经验回归方程

$$\hat{y} = 14.516x_1 + 240.464x_2 - 6.831 \times 10^3$$

10.6 逻辑斯蒂回归分析

10.6.1 逻辑斯蒂回归分析概述

在实际问题中, 还会遇到因变量是定性变量的情况。例如, 在对影响篮球比赛胜负的多因素研究中, 作为因变量篮球比赛的结果只有两种可能, 要么是胜, 要么是负, 它是定性的二分变量。又如, 对影响大学毕业去向的多因素研究中, 如果, 因变量毕业去向共有 6 个可供选择的选项, 则因变量共有 6 种可能的结果, 它同样也是定性的多分类变量。

对这类问题做回归时, 同前面提到的线性回归分析是有区别的, 显然需要一种新的回归方法来替而代之。而逻辑斯蒂回归分析正是用于对定性变量的一种回归分析方法。它分为二元逻辑斯蒂回归和多元逻辑斯蒂回归两种。

当因变量或结果变量是二分变量时, 使用二元逻辑斯蒂回归, 当因变量或结果变量是两类以上的分类变量时, 使用多元逻辑斯蒂回归。

逻辑斯蒂回归仅有很少的条件约束, ①它只须观察是独立的且自变量必须与因变量的对数呈线性关系, ②大样本, 以及, ③同多元线性回归一样要避免自变量间的多重共线性即可。

10.6.2 二元逻辑斯蒂回归分析

1. 模型

设因变量 Y 仅有两个状态, 它们分别以 0 和 1 两个值表示, 我们所要研究的对象是 $p = P(Y=1)$ 。

设有 k 个因素 x_1, x_2, \dots, x_k 影响 Y 的取值, 则逻辑斯蒂线性回归模型为

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

其中, $\beta_0, \beta_1, \dots, \beta_k$ 是待估的未知参数。

由此可得

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}$$

其中, $\frac{p}{1-p}$ 被称为优势比。因此, 可得 p 的计算公式为

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

故, 要计算得到 p 的值, 关键是要对上述的待估参数, 作出估计。

2. 参数估计

设 y 是 0—1 型二分变量, x_1, x_2, \dots, x_k 是与 y 相关的可以精确测定的变量, 实测的 n 组观测数据为 $(x_{i1}, x_{i2}, \dots, x_{ik}; y_i) (i=1, 2, \dots, n)$, 则 y_i 与 $x_{i1}, x_{i2}, \dots, x_{ik}$ 的关系可以表达为

$$E(y_i) = p_i = \beta_0 + \beta_1 + \cdots + \beta_k x_{ik}$$

对于 Logistic 回归

$$f(p_i) = \frac{e^{p_i}}{1 + e^{p_i}} = \frac{e^{\beta_0 + \beta_1 + \cdots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 + \cdots + \beta_k x_{ik}}}$$

y_i 的概率函数为

$$P(y_i) = f(p_i)^{y_i} [1 - f(p_i)]^{1-y_i}$$

$$y_i = 0, 1; i = 1, 2, \dots, n。$$

于是 y_1, y_2, \dots, y_n 的似然函数为

$$l = \prod_{i=1}^n P(y_i) = \prod_{i=1}^n f(p_i)^{y_i} [1 - f(p_i)]^{1-y_i}$$

因而其自然对数函数为

$$L = \ln(l) = \sum_{i=1}^n \{y_i f(p_i) + (1 - y_i)[1 - f(p_i)]\}$$

也就是

$$L = \ln(l) = \sum_{i=1}^n \left[y_i (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) - \ln(1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}}) \right]$$

在计算机上, 通过迭代算法, 可以获得 $\beta_0, \beta_1, \dots, \beta_k$ 的最大似然估计 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ 。

3. 回归系数检验

一个变量要包含在模型中, 它前面的系数应是显著的, 因此, 必须对包含在模型中的变量的回归系数进行显著性检验。

检验变量的回归系数的显著性时, 所要做的原假设

$$H_0: \beta_i = 0 (i=1, 2, \dots, k), \text{ 被择假设为 } H_1: \beta_i \neq 0。$$

在原假设为真时, 检验用的统计量

$$Wald_i = \frac{\hat{\beta}_i}{\text{Var}(\hat{\beta}_i)} \sim \chi^2(k)$$

当 $P(H_0)$ 小于事先给定的显著性水平 α 时, 拒绝原假设, 认为该变量对预测因变量的变化是起作用的。

4. 模型显著性检验

在模型的拟合优度的检验 (即模型的显著性检验) 中, 通常采用如下分类表来反映拟合效果, 见表 10-86。

表 10-86 对因变量的分类表

z		预测值		
		0	1	正确分类比例
观测值	0	n_{00}	n_{01}	f_0
	1	n_{10}	n_{11}	f_1
	总计			f

其中, n_{ij} ($i=0,1; j=0,1$) 表示样本中因变量实际观察值为 i , 而预测值为 j 的样本数

$$f_0 = \frac{n_{00}}{n_{00} + n_{01}} \times 100\% ;$$

$$f_1 = \frac{n_{11}}{n_{10} + n_{11}} \times 100\% ;$$

$$f = \frac{n_{00} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}} \times 100\% .$$

拟合优度的统计量为

$$\sum_{i=1}^n \frac{w_i (y_i - \hat{\pi}_i)^2}{\hat{\pi}_i (1 - \hat{\pi}_i)}$$

式中, n 为总的样本量, w_i 为第 i 个样品的权重, y_i 为两分因变量第 i 个样品的观测值, $\hat{\pi}_i = f(p_i)$ 为第 i 个样品的逻辑斯蒂回归的预测概率。

(1) Hosmer-Lemeshow 检验

该方法根据模型预测概率的大小将所有观察单位十等分 (通常这样, 但有时根据自变量组合和样本量的情况, 组数可能小于 10), 然后根据每一组中因变量的实测值与期望值来计算 Pearson 卡方。

Hosmer-Lemeshow 拟合优度统计量为

$$\chi_{HL}^2 = \sum_{k=1}^g \frac{(O_{1k} - E_{1k})^2}{E_{1k} \xi_K} \sim \chi^2(g-2)$$

式中, g 为所分的组数, O_{1k} 表示第 k 组事件 ($Y=1$) 发生的实际观察频数, E_{1k} 表示第 k 组事件 ($Y=1$) 发生的期望频数, ξ_k 表示第 k 组事件 ($Y=1$) 发生的平均概率。

当 $P(H_0)$ 小于事先给定的显著性水平 α 时, 拒绝原假设 $H_0: O_{1k} = E_{1k}$, 认为模型拟合效果不佳。

观察模型对观测量分类预测的准确性只是判别模型好坏的方法之一, 此外还可以用判别模型与样本间的“相似度”的方法来进行判别。

(2) 似然比检验

这也是一种拟合优度检验方法。我们把利用已有的参数得出的观察结果的可能性称为“似然比”。由于似然比在 $[0,1]$ 之间取值, 所以, 对数似然比 (LL) 值的取值范围在 $[0, -\infty]$ 之间。因此, 习惯上用对数似然比的值乘以 -2 来度量模型对数据的拟合度, 记作 $-2LL$ 。它近似服从自由度为 k 的卡方分布 (k 为自变量数)。其计算公式为

$$-2LL = -2\ln(l) = -2\sum_{i=1}^n \left[y_i (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ip}) - \ln(1 + e^{(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ip})}) \right]$$

它反映了在模型中包括所有自变量后的误差, 用于处理因变量无法解释的变量部分的显著性问题。其值越小, 表明模型的拟合度越好。 $-2LL$ 被称为 Deviance (变异性), 记作 D 。

当在根据研究目的建立逻辑斯蒂回归模型后, 再向模型中引入别的变量重新拟合模型时, 可以分别得到新模型的 $-2LL$ 与原模型的 $-2LL$, 两者之差称为似然比统计量, 根据卡方分布可加性原理可知, 该统计量仍然服从卡方分布。若模型良好, 则新引入的变量应对模型没有改善作用, 此时, 似然比统计量应当无统计学意义。反之, 引入新变量后使模型得到了进一步改善。

5. 模型拟合效果的评价

(1) Cox & Snell R^2

其计算公式为

$$R_{CS}^2 = 1 - \left[\frac{l(0)}{l(\hat{\beta})} \right]^{2/n}$$

其中 $l(0)$ 为初始模型的似然比值, 也就是模型中只包含常数项时的似然比值, $l(\hat{\beta})$ 表示当前模型的似然比值, n 为样本量。对其的解释同多元线性回归中的 R^2 。

由于 Cox & Snell R^2 值不可能为 1 (因 $l(0) \neq 0$), 所以, Nagelkerke 于 1991 年修正了 Cox & Snell R^2 统计量, 从而得到了一个新的统计量 Nagelkerke R^2 。

(2) Nagelkerke R^2

它可在 $0 \sim 1$ 之间取值, 其计算公式为

$$R_N^2 = \frac{R_{CS}^2}{\max(R_{CS}^2)}$$

其中, $\max(R_{CS}^2) = 1 - (l(0))^{2/n}$, 该统计量反映的是由回归方程解释的因变量变异的百分比。

(3) 偏差

对于每个观测量, 其偏差 = $-2\log(\text{观测组的预测概率})$ 。偏差越大, 表明建立的模型没有很好地拟合观测数据。在大样本情况下, 其偏差近似服从正态分布。

6. 实例分析

例 10.22 体质指数 BMI (Body Measure Index) = $\frac{\text{体重 (公斤)}}{[\text{身高 (米)}]^2}$ 是世界卫生组织推

荐用来评价体重是否正常的标准。通常 BMI 大于等于 25, 说明某人肥胖。BMI 越大, 表示某人越肥胖。某高校 3983 名参加体检的职工中, 有 388 位肥胖者, 他们患心血管病的体检数据见表 10-87。(数据来自徐进和李桂枝, 肥胖教职工患心血管病情况的调查和分析, 应用概率统计, 2000 年第 2 期, 220~222)。数据文件见 data10-18.sav。试用二元逻辑斯蒂回归对其进行分析。

表 10-87 体质指数与心血管病的发病率的调查情况

是否患心血管病	体质指数					
	25	26	27	28	29	30
是	68	55	66	32	21	25
否	42	38	20	10	7	4
患病率	0.62	0.59	0.77	0.76	0.75	0.86

本例的具体解题步骤如下:

1. 在数据编辑窗口中, 打开数据文件 data10-18.sav, 并用第 2 章中介绍的方法, 以频数作为加权变量对数据文件进行加权处理。

2. 进行二元逻辑斯蒂回归分析。

按 Analyze→Regression→Binary Logistic 顺序, 打开 Logistic Regression 对话框, 见图 10-35。在左侧变量框中, 选择心血管病变量, 单击向右箭头, 将其送入 Dependent 框; 用同样方法分别将体质指数变量送入 Covariate(s) 框中。

单击 Options 按钮, 在弹出的 Logistic Regression 对话框中, 见图 10-36, 在 Statistics and Plots 项中选择 Hosmer_Lemeshow goodness-of-fit 选项, 要求输出回归方程的拟合程度方面的 Hosmer_Lemeshow 拟合优度检验。

单击 Continue 按钮, 返回 Logistic Regression 对话框。

其他保持系统默认选择。

单击 OK 按钮运行, 则在输出窗中, 得到 11 张表, 见表 10-88 至表 10-98。

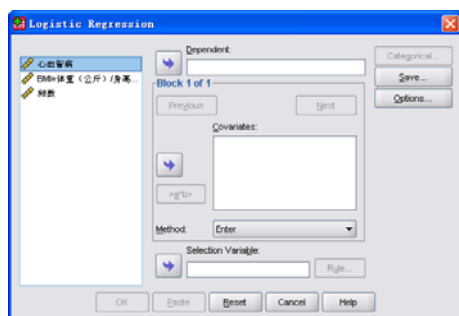


图 10-35 Logistic Regression 对话框

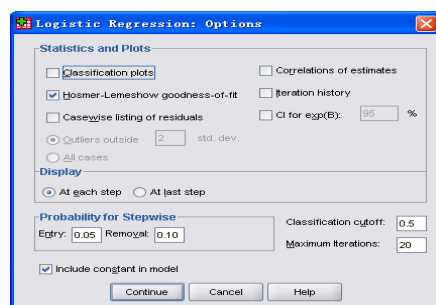


图 10-36 Logistic Regression: Options 对话框

表 10-88 样品处理摘要

Case Processing Summary			
Unweighted Cases ^a		N	Percent
Selected Cases	Included in Analysis	12	100.0
	Missing Cases	0	.0
	Total	12	100.0
Unselected Cases		0	.0
	Total	12	100.0

a. If weight is in effect, see classification table for the total number of cases.

表 10-90 分类表

Classification Table ^{a, b}			
Observed		Predicted	
		心血管病	Percentage Correct
Step 0	否	0	
	是	121	.0
Overall Percentage		0	100.0
		267	68.8

a. Constant is included in the model.

b. The cut value is .500

表 10-92 尚未进入模型的变量

Variables not in the Equation				
Step 0	Variables	Score	df	Sig.
	体质指数	10.925	1	.001
	Overall Statistics	10.925	1	.001

表 10-94 模型摘要

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	470.103 ^a	.029	.041

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

表 10-89 因变量编码

Dependent Variable Encoding

Ori...	Internal Value
否	0
是	1

表 10-91 对初始方程中参数的检验

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.791	.110	52.158	1	.000	2.207

表 10-93 模型系数的综合性检验

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	11.465	1	.001
	Block	11.465	1	.001
	Model	11.465	1	.001

表 10-95 HL 卡方检验

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.886	4	.422

表 10-96 HL 卡方检验的一致性表

Contingency Table for Hosmer and Lemeshow Test					
		心血管病 = 否		心血管病 = 是	
		Observed	Expected	Observed	Expected
Step 1	1	42	44.341	68	65.659
	2	38	31.907	55	61.093
	3	20	24.742	66	61.258
	4	10	9.997	32	32.003
	5	7	5.448	21	22.552
	6	4	4.565	25	24.435
				Total	
				110	
				93	
				86	
				42	
				28	
				29	

表 10-97 分类表

Classification Table ^a			
		Predicted	
		心血管病	
		否	是
Observed			
Step 1	心血管病 否	0	121
	是	0	267
Overall Percentage			68.8

a. The cutvalue is .500

表 10-98 最终在模型中的参数及检验

Variables in the Equation						
		B	S.E.	Wald	df	Sig.
Step 1 ^a	体质指数	.257	.079	10.601	1	.001
	Constant	-6.032	2.089	8.336	1	.004

a. Variable(s) entered on step 1: 体质指数.

3. 结果与讨论。

表 10-88 说明，所有 12 个样品都被有效地用来进行回归分析。

表 10-89 说明因变量的编码情况，原始值否被赋予初始值 0，而原始值是被赋予初始值 1。

表 10-90 显示了在 Step 0 中的拟合情况，可见对于 $y=1$ ，预测的正确率为 100%，而对于 $y=0$ ，预测的正确率为 0%，总的预测正确率为 68.8%。也就是说，如果简单地推测所有人都有心血管病，则意外地将有 68.8 的正确率。

表 10-91 给出了 Step 0 中模型只有常量时对其常量的估计及其检验结果。常量为 0.79146，其标准误为 0.110，Wald 统计量为 52.158，自由度为 1，观测值的显著性水平为 0.000，说明回归模型中的常数项显著， $e^{0.79146} = 2.207$ 。

表 10-92 显示了尚未进入回归模型的变量体质指数的情况，其统计量得分为 10.925，自由度为 1，观测值的显著性水平为 0.001，这意味着，该变量可能在下一步中将进入回归方程。

表 10-93 显示当输入所有自变量（本例只有一个）时整个模型显著（ $P=0.001$ ）。

表 10-94 显示了 Step 1 中模型的摘要，这些值同线性回归分析中的 R^2 相似，后两者给出了用体质指数预测有心血管病方差的一个粗略估计。可见拟合效果不佳。

表 10-95 显示的 HL 卡方检验的结果不拒绝实际观察频数与期望频数相同的原假设（ $P=0.422$ ）。表 10-96 详细地显示了实际观测频数与其期望频数的对照情况，是对表 10-95 的一个补充说明。

表 10-97 是说明第一次迭代结果的拟合效果，可见同表 10-90 一样，说明预测的正确率没有变化。

表 10-98 显示了最终在模型中的变量及其对应回归系数和检验结果。

从 Wald 统计量的检验结果来看, 常数项 ($P=0.004$) 和体质指数的回归系数 ($P=0.001$) 都是显著的, 因此, 可得心血管病发病率的预测方程为

$$\hat{p} = \frac{e^{-6.0323+0.2570 \times BMI}}{1 + e^{-6.0323+0.2570 \times BMI}}$$

而

$$e^B = \frac{p_2 / (1 - p_2)}{p_1 / (1 - p_1)} = e^{0.2570} = 1.293$$

所以, 它就是前面所提到过的比数比 (OR), 在医学上也称为优势之比即优比。由上式转换得到

$$\frac{p_2}{1 - p_2} \approx 1.293 \frac{p_1}{1 - p_1}$$

这说明体质指数 BMI 每升一级, 患心血管病的概率与没有患心血管病的概率的比率 (优势) 将是原来的 1.293 倍。

例 10.23 以例 2.12 中建立的数据文件 data02-17.sav 为基础, 建立不同的年龄和婚姻状态对死亡率的影响的二项逻辑斯蒂回归。

【题析】 用各年龄段的组中值作为各年龄段年龄的代表值, 这样年龄变量可以看作是数值型变量, 而在虚拟变量 m1、m2、m3 中, 未婚状况用 1、0、0 表示, 即 m1=1、m2=0、m3=0, 有配偶状况用 0、1、0 表示, 丧偶状况用 0、0、1 表示, 则离婚状况用 0、0、0 表示。由于要讨论的是死亡率, 故用 1 表示死亡, 用 0 表示健在。

因此, 本例在 SPSS 中的具体解题步骤如下:

1. 在数据编辑窗口中, 打开数据文件 data02-17.sav。
2. 进行二元逻辑斯蒂回归分析。

按 Analyze→Regression→Binary Logistic 顺序, 打开 Logistic Regression 对话框, 见图 10-35。在左侧变量框中, 选择 *健在否* 变量, 单击向右箭头, 将其送入 Dependent 框; 用同样方法分别将 *年龄*、*m1*、*m2*、*m3* 变量送入 Covariate (s) 框中。

按 Continue 按钮, 返回 Logistic Regression 对话框。

其他保持系统默认选择。

按 OK 按钮运行, 则在输出窗中, 得到 9 张表。

由于各表的格式已在例 10.22 的结果已经见过, 而对结论有影响的表格基本都在 Step 1 中, 因此, 本例只对出现在 Step 1 中的表 (见表 10-99 至表 10-102) 进行解释, 其他表的解释参见上例。

3. 结果与讨论。

从表 10-99 可见, 二元逻辑斯蒂回归模型显著 ($P=0.000$)。但从表 10-100 的三个拟合优度的统计量的值可见, 拟合效果欠佳。

表 10-99 对模型系数的综合性检验

Omnibus Tests of Model Coefficients				
Step	Step	Chi-square	df	Sig.
1	Step	1.850E5	4	.000
	Block	1.850E5	4	.000
	Model	1.850E5	4	.000

表 10-100 模型摘要

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	752181.988 ^a	.020	.205

a. Estimation terminated at iteration number 10 because parameter estimates changed by less than .001.

表 10-101 显示了在 Step 1 中的拟合情况, 可见对于 $y=0$ (健在), 预测的正确率为 100%, 而对于 $y=1$ (死亡), 预测的正确率为 0%, 总的预测正确率为 99.1%。也就是说, 如果简单地推测所有人都健在, 则意外地将有 99.1 的正确率。

表 10-101 分类表

Classification Table ^a				
		Predicted		Percentage Correct
		健在否	死亡	
Observed	健在否	9226836	0	100.0
Step 1	死亡	81770	0	.0
Overall Percentage				99.1

a. The cut value is .500

表 10-102 回归系数及其检验

Variables in the Equation						
Step		B	S.E.	Wald	df	Sig.
1 ^a	年龄	.127	.000	6.867E4	1	.000
	m1	1.018	.041	615.271	1	.000
	m2	-.122	.034	13.089	1	.000
	m3	.366	.034	116.593	1	.000
	Constant	-11.979	.046	6.862E4	1	.000

a. Variable(s) entered on step 1: 年龄, m1, m2, m3.

表 10-102 显示了最终在模型中的变量及其对应回归系数和检验结果。

从 Wald 统计量的检验结果来看, 常数项 ($P=0.046$)、年龄和 m1、m2、m3 对应的回归系数 ($P=0.000$ 、0.041、0.034、0.034) 都是显著的, 因此, 可得死亡率关于年龄和婚姻状况的逻辑斯蒂回归的预测方程为

$$\text{未婚状况: } \hat{p} = \frac{e^{-11.979+0.127x+1.018}}{1+e^{-11.979+0.127x+1.018}} = \frac{e^{-10.961+0.127x}}{1+e^{-10.961+0.127x}}$$

$$\text{有配偶状况: } \hat{p} = \frac{e^{-11.979+0.127x-0.122}}{1+e^{-11.979+0.127x-0.122}} = \frac{e^{-12.101+0.127x}}{1+e^{-12.101+0.127x}}$$

$$\text{丧偶状况: } \hat{p} = \frac{e^{-11.979+0.127x+0.366}}{1+e^{-11.979+0.127x+0.366}} = \frac{e^{-11.613+0.127x}}{1+e^{-11.613+0.127x}}$$

$$\text{离婚状况: } \hat{p} = \frac{e^{-11.979+0.127x}}{1+e^{-11.979+0.127x}}$$

10.6.3 多项逻辑斯蒂回归分析

在因变量有 j (大于 2) 个状态, 也即有 j 个水平时, 对它进行逻辑斯蒂回归, 所采用的模型同二分类的逻辑斯蒂回归模型是不一样的。它要通过广义的 Logit 模型的方法来实现。通常, 多项逻辑斯蒂回归模型分两种: 一是其因变量为名义定性变量时的回归模型, 另一个是因变量为有序定性变量时的回归模型。这两种模型的处理方法也是不一样的。

设因变量 y 是 j 类的分类变量, 其各类的赋值依次为 $1, 2, \dots, j$, x_1, x_2, \dots, x_k 是与 y 相关的可以精确测定的变量, 实测的 n 组观测数据为 $(x_{i1}, x_{i2}, \dots, x_{ik}; y_i) (i=1, 2, \dots, n)$, 其中, y_1, y_2, \dots, y_n 为取值 $1, 2, \dots$, 或 j 的随机变量。

1. 名义定性变量的多项逻辑斯蒂回归分析

(1) 模型

在因变量为名义定性变量时, 用其中的一个水平做对照水平 (如以最后一个水平为参照水平), 用每一个分类与对照水平作比较, 就可用来拟合 j 个广义 Logit 模型

$$\begin{cases} \log it p_1 = \ln \left(\frac{p_1}{p_j} \right) = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \dots + \beta_{1k}x_k \\ \log it p_2 = \ln \left(\frac{p_2}{p_j} \right) = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \dots + \beta_{2k}x_k \\ \dots \\ \log it p_j = \ln \left(\frac{p_j}{p_j} \right) = \ln 1 = 0 \end{cases}$$

同时应当有 $p_1 + p_2 + \dots + p_j = 1$ 。其中, $\beta_{10}, \beta_{11}, \dots, \beta_{1k}$ 是待估的未知参数。在计算机上, 通过迭代算法, 可以获得 $\beta_0, \beta_1, \dots, \beta_k$ 的最大似然估计 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ 。

由于每一个分类与对照水平 j 作比较, 故, 如果希望其他任意的两个水平作比较, 只需将两个水平对应的 Logistic 函数相减即可得到相应的函数。

(2) 回归系数检验

同 10.6.2 中一样, 可用 Wald 统计量进行回归系数的显著性检验。

(3) 模型显著性检验

同 10.6.2 中一样, 可用似然比检验。

(4) 模型拟合效果的评价

除用上一节中介绍的 Cox & Snell R^2 、Nagelkerke R^2 外, 还可使用 McFadden 统计量。McFadden 统计量的计算公式为

$$R_{McFadden}^2 = \frac{l(0) - l(\hat{\beta})}{l(0)}$$

其中 $l(0)$ 为初始模型的似然比值, 也就是模型中只包含常数项时的似然比值, $l(\hat{\beta})$ 表示当前模型的似然比值。评判方法同 Cox & Snell R^2 、Nagelkerke R^2 。

(5) 实例分析

例 10.24 研究三个不同学校和两个不同课程计划对学生偏好何种学习方式的影响, 其因变量为学习方式, 收集到的数据资料见表 10-103, 它已存放在数据文件 data10-19.sav 中。试对其进行多项逻辑斯蒂回归分析。

表 10-103 学校、课程计划和学习方式

学校 (x_1)	课程计划 (x_2)	学习方式		
		自学 (1)	小组讨论 (2)	上课 (3)
1	附加 (0)	5	12	50
	常规 (1)	10	17	26
2	附加 (0)	16	12	36
	常规 (1)	21	17	26
3	附加 (0)	12	12	20
	常规 (1)	15	15	16

注：表中学校下方的数字即括号中的数字为数据文件中的值标签，而 x_1 、 x_2 为学校 and 课程计划在回归方程中的变量名。

在 SPSS 中的具体解题步骤如下：

1. 在数据编辑窗口中，打开数据文件 data10-19.sav。
2. 进行多项逻辑斯蒂回归分析。

按 Analyze → Regression → Multinomial Logistic 顺序，打开 Multinomial Logistic Regression 对话框，见图 10-37。在左侧变量框中，选择学习方式变量，单击向右箭头，将其送入 Dependent 框；用同样方法分别将学校、课程计划变量送入 Factor (s) 框中，系统会自动将它们生成虚拟变量。

单击 Statistics 按钮，弹出 Multinomial Logistic Regression: Statistics 对话框，见图 10-38。除保持系统默认选项外，再选择 Goodness-of-fit，要求做模型的拟合优度检验，及选择 Cell Probabilities 和 Classification table 选项，要求计算单元概率和输出分类表。

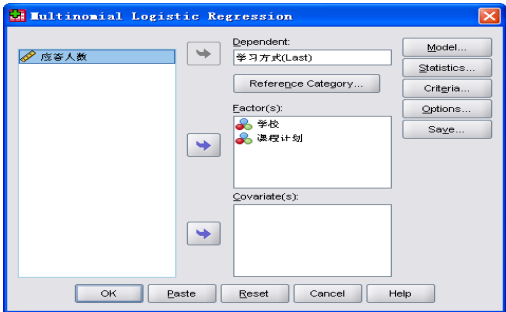


图 10-37 Multinomial Logistic Regression 对话框

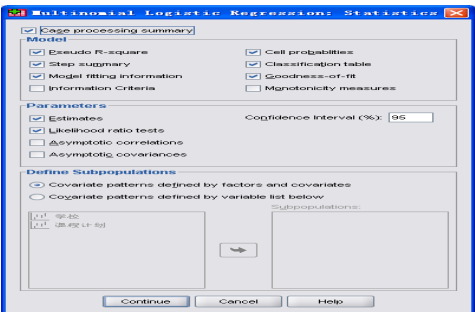


图 10-38 Multinomial Logistic Regression: Statistics 对话框

单击 Continue 按钮，返回 Multinomial Logistic Regression 对话框。

其他保持系统默认选择。

单击 OK 按钮运行，则在输出窗中，得到 8 张表，分别见表 10-104 至表 10-111。

3. 结果与讨论。

表 10-104 显示了样品处理汇总，包括因变量、自变量的分类情况及每一类的观测数及其百分比，同时也包括有效观测数和缺失数据情况，从表中可见本例中没有缺失数据。

表 10-105 显示了最终模型的拟合信息，从表中可见，含有自变量的最终模型和只含有常数项的无效模型相比，Deviance 从 78.195 下降到 50.445，似然比卡方值为 28.470，自由度为 6，似然比卡方检验结果 $P < 0.01$ ，说明至少有一个回归系数不为 0，表明回归模型中至少包含一个自变量，即模型显著。

表 10-104 样品处理汇总

Case Processing Summary			
		N	Marginal Percentage
学习方式	自学	79	23.4%
	小组讨论	85	25.1%
	上课	174	51.5%
学校	学校1	120	35.5%
	学校2	128	37.9%
	学校3	90	26.6%
课程计划	附加	175	51.8%
	正常	163	48.2%
Valid		338	100.0%
Missing		0	
Total		338	
Subpopulation		6	

表 10-105 模型拟合情况

Model Fitting Information				
Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	78.915			
Final	50.445	28.470	6	.000

从表 10-106 可见，模型拟合优度检验不显著 ($P=0.780>0.05$)，表明实际观察数与期望数之间很接近，说明模型可以反映实际情况。

从表 10-107 的伪决定系数来看，三个决定系数的值都很小，不过因为此处只有分类变量，所以它不能给我们表达很多的信息，可以说在此没有多大用处。

表 10-106 模型拟合优度检验

Goodness-of-Fit			
	Chi-Square	df	Sig.
Pearson	1.759	4	.780
Deviance	1.778	4	.777

表 10-107 伪决定系数

Pseudo R-Square	
Cox and Snell	.081
Nagelkerke	.093
McFadden	.041

表 10-108 是分别对每个自变量的作用进行的似然比检验。从表中可见，学校变量对模型的作用在统计学上有极显著性意义 ($P=0.003<0.01$)，课程计划对模型的作用在统计学上也有极显著性意义 ($P=0.004<0.01$)。

表 10-109 显示的是参数估计结果。标识为自学的部分为第一个广义 Logit 模型的参数估计，而标识为小组讨论的部分为第二个广义 Logit 模型的参数估计。其中学校=3 和课程计划=1 为参照。因此，其参数默认为 0，无法估计。

表 10-108 似然比检验

Likelihood Ratio Tests				
Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	50.445 ^a	.000	0	.
学校	66.830	16.385	4	.003
课程计划	61.539	11.094	2	.004

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

表 10-109 参数估计

Parameter Estimates									
		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
学习方式 ^a	Intercept	.091	.292	.098	1	.755	.	.	.
	[学校=1]	-1.315	.384	11.727	1	.001	.269	.127	.570
	[学校=2]	-.232	.333	.486	1	.486	.793	.413	1.522
	[学校=3]	0 ^b	.	.	0
	[课程计划=0]	-.747	.282	7.027	1	.008	.474	.273	.823
	[课程计划=1]	0 ^b	.	.	0
小组讨论	Intercept	.089	.290	.095	1	.758	.	.	.
	[学校=1]	-.656	.339	3.730	1	.053	.519	.267	1.010
	[学校=2]	-.475	.344	1.915	1	.166	.622	.317	1.219
	[学校=3]	0 ^b	.	.	0
	[课程计划=0]	-.743	.271	7.533	1	.006	.476	.280	.809
	[课程计划=1]	0 ^b	.	.	0

a. The reference category is: 上课.

b. This parameter is set to zero because it is redundant.

由此可得, 采用附加教学计划时, 学校 1 中, 学习方式的 Logit 模型的值为

$$g_1 = \ln \frac{p_1}{p_3} = 0.0914 - 1.3147 - 0.7474 = -1.9707$$

$$g_2 = \ln \frac{p_2}{p_3} = 0.0894 - 0.656 - 0.7426 = -1.3092$$

$$g_3 = \ln \frac{p_3}{p_3} = 0$$

每一类的概率值可根据下面的公式计算

$$p(\text{group}_i) = \frac{e^{g_i}}{\sum_{k=1}^j e^{g_k}}$$

由此可以算出采用附加教学计划时, 学校 1 中, 学生偏好各种学习方式概率的预测值为

$$\begin{aligned} p(\text{自学}) &= \frac{e^{-1.7907}}{1 + e^{-1.7907} + e^{-1.3092}} = 0.098874 \\ p(\text{小组讨论}) &= \frac{e^{-1.3092}}{1 + e^{-1.7907} + e^{-1.3092}} = 0.191598 \\ p(\text{上课}) &= \frac{1}{1 + e^{-1.7907} + e^{-1.3092}} = 0.709528 \end{aligned}$$

由此可以计算得到这三种方式下期望人数(由学生偏爱各种学习方式概率的预测值乘以 67)分别为: 6.624558、12.83707 和 47.53838。见表 10-111 学校 1 附加学习计划部分。

按上述的算法也可以依次得到学校 1 正常教学计划下三种学习方式的学习概率, 以及对应的预测人数。实际上, 由此而得的各种条件下的预测人数已列在表 10-93 中。

从表 10-109 可见, 自学与上课两种学习方式相比, 学校 1 的学生比学校 3 的学生更容易选择上课 ($P=0.001$, 优比为 0.269), 而学校 2 与学校 3 的选择上没有统计学上的差异 ($P=0.486$)。按附加学习计划学习的学生比按常规学习计划学习的学生更容易选择上课这一学习方式 ($P=0.008$, 优比为 0.474)。

小组讨论与上课两种学习方式相比, 学校 1、2 与学校 3 的学生选择上, 无统计学上的显著差别 (P 值分别为 0.053 和 0.166)。同样, 按附加学习计划学习的学生比按常规学习计划学习的学生更容易选择上课这一学习方式 ($P=0.006$, 优比为 0.476)。

由此, 可以得到两个模型

$$\begin{aligned} \ln \frac{p_1}{p_3} &= 0.091 - 1.315x_1 - 0.232x_2 - 0.747x_3 \\ \ln \frac{p_2}{p_3} &= 0.089 - 0.656x_1 - 0.475x_2 - 0.743x_3 \end{aligned}$$

式中, x_1 表示第一个学校、 x_2 表示第 2 个学校、 x_3 表示附加学习计划, p_1 为采用自学学习方式的概率, p_2 为采用小组讨论学习方式的概率, p_3 为采用上课学习方式的概率。式中, x_1 与 x_2 的取值为 0, 1, 它们不能同时取 1, $x_1=1$ 时, 表示学校 1, $x_2=1$ 时, 表示学校 2, 而当 $x_1=x_2=0$ 时, 表示学校 3。 x_3 的取值为 0, 1, 取 1 时表示采用附加学习计划, 取 0 时表示采用正常教学计划。

从表 10-110 可见, 偏好自学项上被正确分类的有 15 人, 有 64 人被分到上课项上,

表 10-111 给出了各项上的实际观察频数以及由预测模型给出的各行的预测频数、Pearson 预测残差、观察数占该层总观察数的百分比、预测频数占该层总观察数的百分比。

表 10-111 实际观察频数和预测频数

Classification				
Observed	Predicted			
	自学	小组讨论	上课	Percent Correct
自学	15	0	64	19.0%
小组讨论	15	0	70	.0%
上课	16	0	158	90.8%
Overall Percentage	13.6%	.0%	86.4%	51.2%

Observed and Predicted Frequencies							
			Frequency		Percentage		
学校	课程 附加	学习方式	Observed	Predicted	Pearson Residual	Observed	Predicted
学校1	附加	自学	5	6.624	-.665	7.5%	9.9%
		小组讨论	12	12.841	-.261	17.9%	19.2%
		上课	50	47.535	.663	74.6%	70.9%
	正常	自学	10	8.376	.612	18.9%	15.8%
		小组讨论	17	16.159	.251	32.1%	30.5%
		上课	26	28.465	-.679	49.1%	53.7%
学校2	附加	自学	16	15.180	.241	25.0%	23.7%
		小组讨论	12	11.932	.022	18.8%	18.6%
		上课	36	36.889	-.225	56.2%	57.6%
	正常	自学	21	21.820	-.216	32.8%	34.1%
		小组讨论	17	17.068	-.019	26.6%	26.7%
		上课	26	25.111	.227	40.6%	39.2%
学校3	附加	自学	12	11.196	.278	27.3%	25.4%
		小组讨论	12	11.228	.267	27.3%	25.5%
		上课	20	21.576	-.475	45.5%	49.0%
	正常	自学	15	15.804	-.250	32.6%	34.4%
		小组讨论	15	15.772	-.240	32.6%	34.3%
		上课	16	14.424	.501	34.8%	31.4%

The percentages are based on total observed frequencies in each subpopulation

当定性因变量的各类之间有等级关系且分类数大于 2 时, 称这类资料为多分类有序因变量的资料。如, 运动员等级分为国际健将、健将、一级、二级、三级五个等级、学习成绩分成优、良、中、差四个等级等。对这种资料进行逻辑斯蒂回归时, 需要拟合分类数减 1 个 logit 模型, 此时的模型被称为累加 logit 模型。

设因变量共有 $1, 2, \dots, j$ 个有序类别，自变量有 k 个时，则应当同时拟合 $j-1$ 个累加 logit 模型

$$\left. \begin{aligned} \log it_1 &= \ln \frac{p_1}{1-p_1} = \beta_{10} + \beta_1 x_1 + \cdots + \beta_k x_k \\ \log it_2 &= \ln \frac{p_1 + p_2}{1-p_1-p_2} = \beta_{20} + \beta_1 x_1 + \cdots + \beta_k x_k \\ &\vdots \\ \log it_{j-1} &= \ln \frac{p_1 + p_2 + \cdots + p_{j-1}}{1-p_1-p_2-\cdots-p_{j-1}} = \beta_{(j-1)0} + \beta_1 x_1 + \cdots + \beta_k x_k \end{aligned} \right\}$$

模型中最后一类被作为对比的基础水平, p_1 、 p_2 、 \cdots 、 p_{j-1} 分别为因变量取第一类、第二类、 \cdots 、第 $j-1$ 类时的概率。从模型中不难看出, 这种模型实际上是依次将因变量

划分为两个等级, 另外, 除常数项在不断改变外, 所有自变量的系数是固定不变的, 因此, 此时所求出的优比 (OR) 是自变量每改变一个单位, 因变量提高一个及一个以上等级的比数比。

同前面一样, 对回归系数的求法, 可在计算机上, 通过迭代算法, 来获得 $\beta_0, \beta_1, \dots, \beta_k$ 的最大似然估计 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ 。

对回归系数检验时, 同 10.6.2 中一样, 可用 Wald 统计量进行回归系数的显著性检验。而对进行模型显著性检验时, 同 10.6.2 中一样, 可用似然比检验。

此外, 对模型拟合效果的评价, 也可用上节中提到的 Cox & Snell R^2 、Nagelkerke R^2 和 McFadden 统计量。

例 10.25 试根据表 10-112 所示的 970 个患者的数据, 研究性别、年龄和两种治疗方法对某种疾病疗效的影响。数据已存放在 data10-20.sav 中。

表 10-112 性别、年龄、治疗方法和疗效

性别	年龄	治疗方法	治疗		
			无效	较有效	有效
男	中年	传统疗法	19	23	24
		新疗法	22	42	24
	老年	传统疗法	57	47	37
		新疗法	45	56	55
女	中年	传统疗法	29	33	42
		新疗法	10	27	36
	老年	传统疗法	63	66	68
		新疗法	42	53	50

1. 在数据编辑窗口中, 打开数据文件 data10-20.sav。
2. 进行有序逻辑斯蒂回归分析。

按 Analyze→Regression→Ordinal 顺序, 打开 Ordinal Regression 对话框, 见图 10-39。

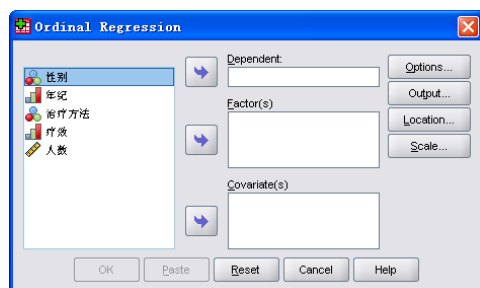


图 10-39 Ordinal Regression 对话框

在左侧变量框中, 选择疗效变量, 单击向右箭头, 将其送入 Dependent 框; 用同样方法分别将年纪、性别、治疗方法变量送入 Factor (s) 框中, 系统会自动将它们生成虚拟变量。

单击 Continue 按钮, 返回 Ordinal Regression 对话框。其他保持系统默认选择。

单击 OK 按钮运行, 则在输出窗中, 得到 5 张表, 分别见表 10-113 至表 10-117。

3. 结果与讨论。

表 10-113 显示了样品处理汇总, 包括因变量、自变量的分类情况及每一类的观测数及其百分比, 同时也包括有效观测数和缺失数据情况, 从表中可见本例中总观察人数为 970 人, 没有缺失数据。

表 10-114 显示了模型拟合情况, 含有自变量的最终模型, 在统计学上有显著性意义 ($\chi^2 = 14.514, p = 0.002$)。

表 10-113 样品处理汇总

Case Processing Summary		
	N	Marginal Percentage
疗效	有效	336 34.6%
	较有效	347 35.8%
	无效	287 29.6%
年纪	中年	331 34.1%
	老年	639 65.9%
性别	女	519 53.5%
	男	451 46.5%
治疗方法	传统疗法	508 52.4%
	新疗法	462 47.6%
Valid	970	100.0%
Missing	0	
Total	970	

表 10-114 模型拟合情况

Model Fitting Information				
Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	106.152			
Final	91.637	14.514	3	.002

Link function: Logit.

表 10-115 显示的 Pearson 和 Deviance 检验结果表明, 模型拟合较好 ($P=0.280$ 和 0.267)。

表 10-116 显示的是伪决定系数, 由于自变量均为分类变量, 因此, 其解释作用不大。

表 10-115 拟合优度检验

Goodness-of-Fit			
	Chi-Square	df	Sig.
Pearson	13.206	11	.280
Deviance	13.418	11	.267

Link function: Logit.

表 10-116 伪决定系数

Pseudo R-Square	
Cox and Snell	.015
Nagelkerke	.017
McFadden	.007

Link function: Logit.

表 10-117 显示的是对模型中的参数所作的估计。共给出 5 个参数, 前两个为常数项, 后三个分别为年纪、性别和治疗方法变量的参数。

从表中可见, 不同年纪、性别和治疗方法对疗效的作用, 在统计学上有显著性意义 ($P=0.016$ 、 0.18 和 0.046)。由此, 可得有序分类资料的逻辑斯蒂回归方程为

$$\begin{cases} \log it_1 = \ln \frac{p_1}{1-p_1} = -0.775 - 0.302x_1 - 0.282x_2 + 0.239x_3 \\ \log it_2 = \ln \frac{p_1 + p_2}{1-p_1-p_2} = 0.745 - 0.302x_1 - 0.282x_2 + 0.239x_3 \end{cases}$$

表 10-117 参数估计

Parameter Estimates								
	Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval		
						Lower Bound	Upper Bound	
Threshold [疗效 = 1]	-.775	.117	43.798	1	.000	-1.005	-.546	
[疗效 = 2]	.745	.117	40.625	1	.000	.516	.974	
Location [年纪=1]	-.302	.125	5.819	1	.016	-.548	-.057	
[年纪=2]	0 ^a	.	.	0	.	.	.	
[性别=0]	-.282	.120	5.552	1	.018	-.517	-.047	
[性别=1]	0 ^a	.	.	0	.	.	.	
[治疗方法=1]	.239	.120	3.980	1	.046	.004	.473	
[治疗方法=2]	0 ^a	.	.	0	.	.	.	

Link function: Logit.

a. This parameter is set to zero because it is redundant.

式中, p_1 表示治疗有效的概率, p_2 表示治疗较有效的概率, x_1 表示年纪 (年纪为中年时取 1, 老年时取 0), x_2 表示性别 (性别为女时取 1, 性别为男时取 0), x_3 表示治疗方法 (治疗方法 1 时取 1, 治疗方法 2 时取 0)。

从上面的公式中, 可以得到, 中年疗效比老年低, 其优比 (OR) 值为 $e^{-0.302} = 0.7394$; 女性疗效比男性低, 其优比 (OR) 值为 $e^{-0.282} = 0.7543$; 新疗法比传统疗法好, 其优比 (OR) 值为 $e^{0.239} = 1.27$ 。

第 11 章 聚类分析与判别分析

古语说得好“物与类聚，人与群分”。分类问题在科学研究中是最常见的。为探索事物发展的内在规律性，我们通常要用类比的方法，加以分析研究，找出它们间的区别和联系。而对同类事物的研究，可得到其内在的共性和个性。因此，要使这些问题得到较好的解决，首先必须解决好分类问题。

对于分类，一般我们都是靠经验和借助于专业知识进行的，很多时候我们都可这样做。但随着人类对自然的认识的不断深入，分类越来越细，要求也越来越高，有时我们光凭经验和专业知识已无法再进行精确分类，尤其是对事物的性质不很明确时，我们就很可能难以下手。此时，我们必须借助于数值分类学的方法，而今已从中分离出来的聚类分析的方法来进行研究。

聚类分析是研究分类问题的一种多元统计方法。对类作出严格的数学定义是一件很麻烦的事情，在不同问题中对类的定义也是各不相同的。这里所谓的类是指相似或相近元素的集合。

所以聚类分析的目的是把相似的或相近的对象归并成类，研究的主要内容是如何度量相似性和构造聚类的具体方法。聚类分析分为两种方法：一种是对测定指标进行分类的叫“指标聚类”，又称“R 型聚类”；另一种是针对样品进行分类的叫“样品聚类”，又称“Q 型聚类”。

判别分析是判别样品所属类型的一种统计方法。它是在已知研究对象分成若干类别，并已取得各种类别的一批已知样品的观测数据的基础上，根据某些准则建立判别式，再对未知类型的样品进行判别分析。

11.1 聚 类 分 析

11.1.1 聚类分析的作用

应用聚类分析主要解决以下几类问题：

对分类不明确的指标进行分类，有助于加深对这类指标内部规律的了解和研究。

通过对指标的分类，可找出各类的典型指标，这样可大大缩减指标量，有助于做进一步的研究。如进行选材研究和综合评价时，一方面可了解所选指标面的广泛性，另一方面还可减少同类相关指标的重复性；在进行多元线性回归分析时，由于自变量的共线

性导致偏回归系数不能真正反映自变量对因变量的影响。因此往往先要进行变量聚类，找出彼此独立且有代表性的自变量，而又不丢失大部分信息，这样才能使所得的回归方程更具实用性。

指标的减少，也可使研究所需的样本含量也能成倍地有效减少，这样可大大提高科研的效益和效率。

对样品的聚类，有助于对不同国家、地区、单位、项目或个体进行分类，从而为构造判别函数做好准备。

11.1.2 聚类分析中常用的统计量

聚类分析中，通常用来分类的统计量有：反映相似性的相似测度（通过计算 Pearson 相关系数或 Cosine 相似系数来表示）和反映不相似性的不相似性测度（距离系数）。

为了将指标或样品进行分类，就需要研究指标或样品之间的关系。目前用得最多的方法有两个：一个是相似系数，性质越接近的指标或样品，它们的相似系数的绝对值越接近于 1，而彼此无关的指标或样品的相似系数的绝对值越接近于 0。比较相似的指标或样品归为一类，不怎么相似的指标或样品归为不同的类。另一种方法是将一个样品或指标看作 P 维空间的一个点，并在空间定义距离，距离越近的点归为一类，距离较远的点归为不同的类。

相似系数和距离系数有各种不同定义，这些定义取决于变量的类型。

在实际问题中，我们将变量按以下三种尺度进行分类：

- 间隔尺度：变量是用连续的量来表示的，如长度、重量、速度、压力等。在间隔尺度中，如果存在绝对零点，又称比例尺度。
- 有序尺度：变量度量时没有明确的数量表示，而是划分一些等级，等级之间有次序关系。如，优、良、中、差等，此四等间没有数量关系，但有次序关系。
- 名义尺度：变量度量时既没有明确的数量关系，也没有次序关系。如，物体的红、黄、白三种颜色，医学上的阳性和阴性等。

不同的变量类型要选择不同的统计量。

具体统计量的选用及其计算公式详见 10.3.1 中的相关内容。

11.1.3 系统聚类法

11.1.3.1 基本思想

首先，把每个变量（每个样品）看作一类，并规定变量间的相似性测度换算成的距离（ $d_{ij}^2 = 1 - c_{ij}^2$ ，其中 c_{ij} 表示变量 i 和变量 j 之间的相关系数，或样品 i 和样品 j 之间的相似系数）（或样品之间的距离）看作类与类之间的距离，然后将距离最近的两类合成新的一类，每次减少一类，重新计算新类与其他各类的距离，重复进行最近类的合并，直

至所有的变量（或样品）合并成一类。

11.1.3.2 方法

类与类之间的距离的定义如同样品间的距离定义一样，有各种各样不同的方法。类与类之间用各种不同的方法定义的距离就产生了各种不同的聚类分析的方法。常用的聚类分析的方法有系统聚类、动态聚类等的方法。其中，系统聚类方法是用得最多的一种方法。

设 d_{ij} 表示样品 i 与样品 j 之间的距离， G_1, G_2, \dots 表示类， D_{ij} 表示 G_i 与 G_j 的距离。

在有关距离的系统聚类中常用的方法有八种，它们分别是：

1. 最短距离法

定义类 G_i 与 G_j 之间的距离为两类最近样品（或指标）的距离，即

$$D_{ij} = \min_{x_i \in G_i, x_j \in G_j} d_{ij}$$

设 G_p 与 G_q 合并成一个新类，记为 G_r ，则任一类 G_k 与 G_r 的距离是

$$D_{kr} = \min_{x_i \in G_i, x_j \in G_j} d_{ij} = \min\{D_{kp}, D_{kq}\}$$

最短距离法聚类的步骤如下：

(1) 定义样品（变量）之间的距离，计算 n 个样品（变量）间的距离，得到距离矩阵 $D_{(0)}$ ，它是一个对称矩阵。开始每个样品（变量）自成一类，显然，此时 $D_{ij} = d_{ij}$ 。

(2) 找出 $D_{(0)}$ 中非对角线中的最小元素，设它为 D_{pq} ，则将 G_p 与 G_q 合并成一个新类，记为 G_r ，即 $G_r = \{G_p, G_q\}$ 。在 $D_{(0)}$ 中划去 G_p 与 G_q 对应的两行和两列。

(3) 计算新类 G_r 与任一类 G_k 的距离

$$D_{kr} = \min_{x_i \in G_i, x_j \in G_j} d_{ij} = \min\{D_{kp}, D_{kq}\}$$

将计算得到的新类 G_r 与剩下的未聚类的各类之间的距离所组成的一行和一列添加到 $D_{(0)}$ 中，其余行和列上的距离值不变，这样可形成 $D_{(1)}$ 。

(4) 对 $D_{(1)}$ 重复上述对 $D_{(0)}$ 的(2)、(3)两步得 $D_{(2)}$ ，像这样循环下去，直至所有元素合并成一类为止。

如果在上述任意一步中， $D_{(k)}$ 非对角线上的最小元素不止一个，称这种现象为结，对应这些最小元素的类可以同时对其合并成一类。

2. 最长距离法

定义类 G_i 与 G_j 之间的距离为两类最远样品（或指标）的距离，即

$$D_{ij} = \max_{x_i \in G_i, x_j \in G_j} d_{ij}$$

最长距离法聚类的步骤同最短距离法的聚类步骤类似。当 G_p 与 G_q 合并成一个新类

(记为 G_r) 时, 则任一类 G_k 与 G_r 的距离的计算公式为

$$D_{kr} = \max_{x_i \in G_i, x_j \in G_j} d_{ij} = \max\{D_{kp}, D_{kq}\}$$

3. 中间距离法

定义类与类之间的距离既不采用最近距离, 也不采用最远距离, 而是采用两类之间的距离, 故称为中间距离法。它的聚类的步骤同最短距离法的聚类步骤类似, 当 G_p 与 G_q 合并成一个新类 (记为 G_r) 时, 则任一类 G_k 与 G_r 的距离的计算公式为

$$D_{kr}^2 = \frac{1}{2} D_{kp}^2 + \frac{1}{2} D_{kq}^2 + \beta D_{pq}^2$$

其中 $-\frac{1}{4} \leq \beta \leq 0$ 。

4. 重心法

它的聚类的步骤同最短距离法的聚类步骤类似, 当 G_p 与 G_q 合并成一个新类时 (记为 G_r), 则任一类 G_k 与 G_r 的距离的计算公式为

$$D_{kr}^2 = \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2 - \frac{n_p}{n_r} \frac{n_q}{n_r} D_{pq}^2$$

式中, n_p 、 n_q 、 n_r 分别表示第 p 、 q 、 r 类中的样品数。

5. 类平均法:

定义两类之间的距离平方为这两类元素两两之间距离平方的平均, 即

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{x_i \in G_i} \sum_{x_j \in G_j} d_{ij}^2$$

式中, n_p 、 n_q 分别表示第 p 、 q 类中的样品数。

它的聚类的步骤同最短距离法的聚类步骤类似, 当 G_p 与 G_q 合并成一个新类 (记为 G_r) 时, 则任一类 G_k 与 G_r 的距离的计算公式为

$$D_{kr}^2 = \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2$$

式中, n_p 、 n_q 、 n_r 分别表示第 p 、 q 、 r 类中的样品数。

6. 可变类平均法

定义两类之间的距离平方为这两类元素两两之间距离平方的平均, 即

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{x_i \in G_i} \sum_{x_j \in G_j} d_{ij}^2$$

式中, n_p 、 n_q 分别表示第 p 、 q 类中的样品数。

它的聚类的步骤同最短距离法的聚类步骤类似, 当 G_p 与 G_q 合并成一个新类 (记为

G_r) 时, 则任一类 G_k 与 G_r 的距离是

$$D_{kr}^2 = \frac{n_p}{n_r}(1-\beta)D_{kp}^2 + (1-\beta)\frac{n_q}{n_r}D_{kq}^2 + \beta D_{pq}^2$$

式中, n_p 、 n_q 、 n_r 分别表示第 p 、 q 、 r 类中的样品数。其中 β 是可变的, 且 $\beta < 1$ 。

7. 可变法

定义两类之间的距离平方为这两类元素两两之间距离平方的平均, 即

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{x_i \in G_i} \sum_{x_j \in G_j} d_{ij}^2$$

式中, n_p 、 n_q 分别表示第 p 、 q 类中的样品数。

它的聚类的步骤同最短距离法的聚类步骤类似, 当 G_p 与 G_q 合并成一个新类 (记为 G_r) 时, 则任一类 G_k 与 G_r 的距离计算公式为

$$D_{kr}^2 = \frac{1-\beta}{2}(D_{kp}^2 + D_{kq}^2) + \beta D_{pq}^2$$

其中 β 是可变的, 且 $\beta < 1$ 。

8. 离差平方和法

它的聚类的步骤同最短距离法的聚类步骤类似, 当 G_p 与 G_q 合并成一个新类 (记为 G_r) 时, 则任一类 G_k 与 G_r 的距离的计算公式为

$$D_{kr}^2 = \frac{n_k + n_p}{n_r + n_k} D_{kp}^2 + \frac{n_k + n_q}{n_r + n_q} D_{kq}^2 - \frac{n_k}{n_r + n_k} D_{pq}^2$$

式中, n_k 、 n_p 、 n_q 、 n_r 分别表示第 k 、 p 、 q 、 r 类中的样品数。

11.1.3.3 实例分析

1. 指标聚类 (R 型聚类)

例 11.1 1984 年洛杉矶奥运会发布了 55 个国家的男子径赛纪录, 见表 11-1, 数据已存放在 data11-01.sav 中, 试对 100 米、200 米、400 米、800 米、1500 米、5000 米、10000 米、马拉松 8 个指标作聚类分析。

在 SPSS 中进行类似本例作指标聚类的步骤如下:

1. 在数据编辑窗口中, 打开 data11-01.sav。

2. 按 Analyze→Classify→Hierarchical Cluster 顺序, 打开 Hierarchical Cluster Analysis (分层聚类, 也就是常说的系统聚类) 对话框, 见图 11-1。

3. 在左侧的变量名框中, 同时选择 @100 米、@200 米、@400 米、@800 米、@1500 米、@5000 米、@10000 米、马拉松变量, 单击右移箭头按钮, 将它们移入 Variable(s) 矩形框中。

表 11-1 1984 年 55 个国家或地区男子径赛纪录表

	国家或地区	@100米	@200米	@400米	@800米	@1500米	@5000米	@10000米	马拉松
1	Argentina	10.39	20.81	46.84	1.81	3.70	14.04	29.36	137.72
2	Australia	10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.30
3	Austria	10.44	20.81	46.82	1.79	3.60	13.26	27.72	135.90
4	Belgium	10.34	20.68	45.04	1.73	3.60	13.22	27.45	129.95
5	Bermude	10.28	20.58	45.91	1.80	3.75	14.68	30.55	146.62
6	Brazil	10.22	20.43	45.21	1.73	3.66	13.62	28.62	133.13
7	Burma	10.64	21.52	48.30	1.80	3.85	14.45	30.28	139.95
8	Canada	10.17	20.22	45.68	1.76	3.63	13.55	28.09	130.15
9	Chile	10.34	20.80	46.20	1.79	3.71	13.61	29.30	134.03
10	China	10.51	21.04	47.30	1.81	3.73	13.90	29.13	133.53
11	Columbia	10.43	21.05	46.10	1.82	3.74	13.49	27.88	131.35
12	Cookis	12.18	23.20	52.94	2.02	4.24	16.70	35.38	164.70
13	Costa	10.94	21.90	48.66	1.87	3.84	14.03	28.81	136.58
14	Czech	10.35	20.65	45.64	1.76	3.58	13.42	28.19	134.32
15	Denmark	10.56	20.52	45.89	1.78	3.61	13.50	28.11	130.78
16	Domrep	10.14	20.65	46.80	1.82	3.82	14.91	31.45	154.12
17	Finland	10.43	20.69	45.49	1.74	3.61	13.27	27.52	130.87
18	France	10.11	20.38	45.28	1.73	3.57	13.34	27.97	132.30
19	Gdr	10.12	20.33	44.87	1.73	3.56	13.17	27.42	129.92
20	Frg	10.16	20.37	44.50	1.73	3.53	13.21	27.61	132.23
21	Gbni	10.11	20.21	44.93	1.70	3.51	13.01	27.51	129.13
22	Greece	10.22	20.71	46.56	1.78	3.64	14.59	28.45	134.60
23	Guatemala	10.98	21.82	48.40	1.89	3.80	14.16	30.11	139.33
24	Hungary	10.26	20.62	46.02	1.77	3.62	13.49	28.44	132.58
25	India	10.60	21.42	45.73	1.76	3.73	13.77	28.81	131.98
26	Indonesia	10.59	21.49	47.80	1.84	3.92	14.73	30.79	148.83
27	Ireland	10.61	20.96	46.30	1.79	3.56	13.32	27.81	132.35
28	Israel	10.71	21.00	47.80	1.77	3.72	13.66	28.93	137.55
29	Italy	10.01	19.72	45.26	1.73	3.60	13.23	27.52	131.08
30	Japan	10.34	20.81	45.86	1.79	3.64	13.41	27.72	128.63
31	Kenya	10.46	20.66	44.92	1.73	3.55	13.10	27.38	129.75
32	Korea	10.34	20.89	46.90	1.79	3.77	13.96	29.23	136.25
33	Dprkorea	10.91	21.94	47.30	1.85	3.77	14.13	29.67	130.87
34	Luxembou	10.35	20.77	47.40	1.82	3.67	13.64	29.08	141.27
35	Malaysia	10.40	20.92	46.30	1.82	3.80	14.64	31.01	154.10
36	Mauritlu	11.19	22.45	47.70	1.88	3.83	15.06	31.77	152.23
37	Mexico	10.42	21.30	46.10	1.80	3.65	13.46	27.95	129.20
38	Netherla	10.52	20.95	45.10	1.74	3.62	13.36	27.61	129.02
39	Nz	10.51	20.88	46.10	1.74	3.54	13.21	27.70	128.98
40	Norway	10.55	21.16	46.71	1.76	3.62	13.34	27.69	131.48
41	Png	10.96	21.78	47.90	1.90	4.01	14.72	31.36	148.22
42	Philippi	10.78	21.64	46.24	1.81	3.83	14.74	30.64	145.27
43	Poland	10.16	20.24	45.36	1.76	3.60	13.29	27.89	131.58
44	Portugal	10.53	21.17	46.70	1.79	3.62	13.13	27.38	128.65
45	Rumania	10.41	20.98	45.87	1.76	3.64	13.25	27.67	132.50
46	Singapor	10.38	21.28	47.40	1.88	3.89	15.11	31.32	157.77
47	Spain	10.42	20.77	45.98	1.76	3.55	13.31	27.73	131.57
48	Sweden	10.25	20.61	45.63	1.77	3.61	13.29	27.94	130.63
49	Switzerl	10.37	20.46	45.78	1.78	3.55	13.22	27.91	131.20
50	Taipei	10.59	21.29	46.80	1.79	3.77	14.07	30.07	139.27
51	Thailand	10.39	21.09	47.91	1.83	3.84	15.23	32.56	149.90
52	Turkey	10.71	21.43	47.60	1.79	3.67	13.66	28.58	131.50
53	USA	9.93	19.75	43.86	1.73	3.53	13.20	27.43	128.22
54	USSR	10.07	20.00	44.60	1.75	3.59	13.20	27.53	130.55
55	Wsamoa	10.82	21.86	49.00	2.02	4.24	16.28	34.71	161.83

由于本例要做指标（变量）聚类，因此，在 Cluster 栏中选择 Variables，要求进行 R 型指标聚类。

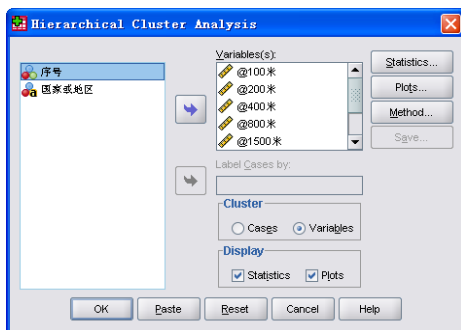


图 11-1 Hierarchical Cluster Analysis 对话框

在 Display 栏中，有两项选项，必须选择一项，本例中选择 Plots，主要考虑分类时用图示比较直观。

4. 单击 Plots 按钮，打开 Hierarchical Cluster Analysis Plots 对话框，见图 11-2。选择 Dendrogram，要求输出树形图。为适合不同研究者对图示有不同的爱好，本例中还选择 Icicle（冰柱图形）多项选择

中的 All Clusters 选项, 要求将聚类中的每一步都体现在图中, 需要指出的是此时的变量还不算太多, 否则应选用第二个选项, 自己定义分类数。

在 Orientation (显示方向) 中, 选择 Horizontal 选项 (水平显示), 如果希望垂直显示, 则可选第一项 Vertical。

单击 Continue 按钮, 返回 Hierarchical Cluster Analysis 对话框 (见图 11-1)。

5. 单击 Method 按钮, 打开 Hierarchical Cluster Analysis: Method 对话框, 见图 11-3。

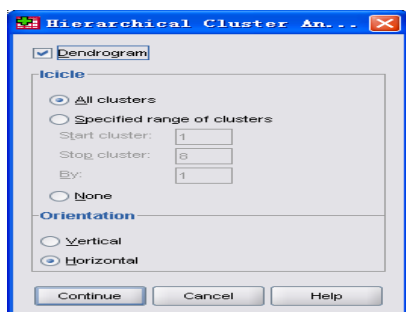


图 11-2 Hierarchical Cluster Analysis: Plots 对话框

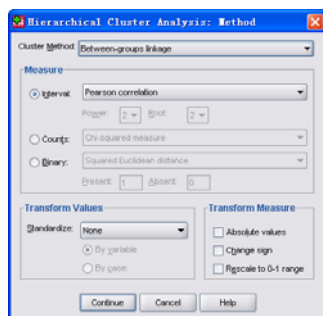


图 11-3 Hierarchical Cluster Analysis: Method 对话框

在 Cluster Method (聚类方法) 的下拉列表中, 共有 7 中聚类方法, 见图 11-4。分别为: 类平均法、组内联结法、最短距离、最长距离、重心法、中间距离法和 Ward 最小方差法。其中类平均法被实践证明是十分稳健的方法, 故一般研究中都采用此法。因此, 本例选择系统的默认选择项 Between-groups linkage (类平均法)。

在 Measure 中, 选择距离或相似性测度的方法。

本例的数据是尺度类型, 因此, 只能选择第一项 Interval。在 Interval 中, 具体有八项计算距离的方法, 见图 11-5, 分别是:

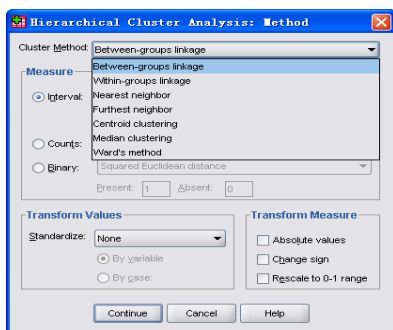


图 11-4 聚类方法

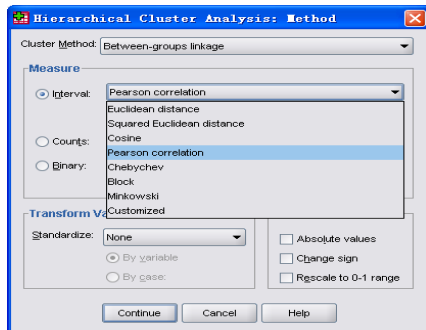


图 11-5 Interval 列表中的选项

(1) Euclidean distance 欧几米德 (欧氏) 距离。

(2) Squared Euclidean distance 欧氏距离平方。

(3) Cosine cos 相似性测度。

(4) Pearson correlation 皮尔逊相关。

(5) Chebychev 切贝谢夫距离。

(6) Block 布洛克距离。

(7) Minkowski 明可斯基距离。

上述这些方法的计算公式详见 10.3.1 中的相关内容。

(8) Customized 选项, 自定义距离, 两项之间的距离用各项值之间差值绝对值的 p 次幂之和的 r 次方根表示。计算公式为

$$MINKOSKI(x, y) = \sqrt[r]{\sum_{i=1}^n |x_i - y_i|^p}$$

式中, n 为样本含量。

由于在计量资料的指标聚类中, 一般采用皮尔逊相关。故本例选择 Pearson correlation。

在 Transform Values 选项中, 可在 Standardize 标准化方法的下拉列表中选择, 见图 11-6。

对数据进行标准化的方法有:

(1) 系统默认项 None, 即对原始资料不做任何处理。

(2) Z scores, 把数值标准化到 Z 分数。标准化后变量的均值为 0, 标准差为 1。

$$x'_i = \frac{x_i - \bar{X}}{S}$$
 如果原始数据的标准差为 0 (即原始数据为常量), 则将所有值置为 0。

(3) Range -1 to 1, 把数值标准化到 -1 至 1 范围内。选择该项, 对每个值用正在被标准化的变量或观测量的值除以其两极差。如果两极差为 0, 所有值不变。

(4) Maximum magnitude of 1, 把数值标准化到最大值为 1。该方法是把正在标准化的变量或观测量的值用最大值去除。如果最大值为 0, 则用最小值的绝对值除再加 1。

(5) Range 0 to 1 选项, 把数值标准化到 0 至 1 的范围内, 对正在被标准化的变量或观测量的值减去其最小值, 然后除以两极差。如果两极差是 0, 将所有变量值或观测测量值设置为 0.5。

(6) Mean of 1 选项, 把数值标准化到以均值为单位。对正在被标准化的变量或观测量的值除以其均值。如果均值是 0, 对变量或观测量的所有值都加 1, 使其均值为 1。

(7) Standard deviation of 1 选项, 把数值标准化到以标准差为单位。该方法对每个值除以其标准差。如果标准差为 0, 则这些值保持不变。

考虑到聚类过程中, 新变量合并时最好使得变量单位量纲之间保持一致, 故习惯上

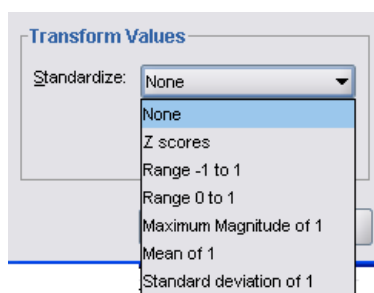


图 11-6 数据资料标准化方法

选择 Z scores。

其他选项采用系统默认值，单击 Continue 按钮，返回 Hierarchical Cluster Analysis 对话框。

6. 单击 OK 按钮执行，则在输出窗口中，得到如图 11-7 和图 11-8 所示的聚类图。

Horizontal Icicle							
Case	Number of clusters						
	1	2	3	4	5	6	7
马拉松	×	×	×	×	×	×	×
@10000米	×	×	×	×	×	×	
@5000米	×	×	×	×	×	×	×
@1500米	×	×	×	×	×	×	×
@800米	×	×	×	×	×	×	×
@400米	×	×	×	×	×	×	×
@200米	×	×	×	×	×	×	×
@100米	×	×	×	×	×	×	×

图 11-7 聚类冰柱图

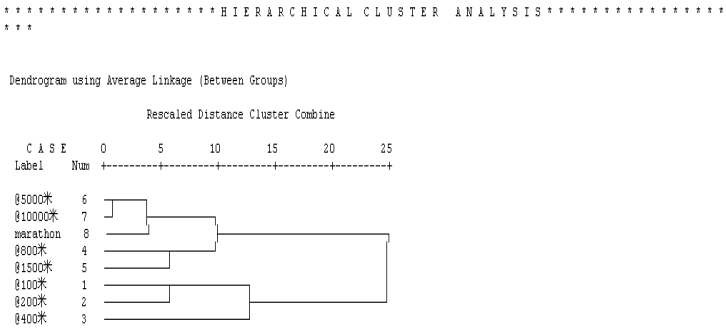


图 11-8 聚类树状图

7. 结果与分析。

从冰柱图可见，聚类第一步八个变量全部成一类，第二步八个变量分成 2 类，看×号断开处，作为分界，800 米以上为一类，400 米以下为一类，余类推，可见八个变量共分 4 类比较合适，分别为 100 米、200 米一类，400 米一类，800 米、1500 米一类及 5000 米、10000 米、马拉松一类。

从树状图可见，在 7 左右将树状断开（究竟从何处断开，要结合专业知识而定），可将八个指标变量分为 4 类，同上。

8. 分类命名。

100 米、200 米短距离类，400 米中距离类、800 米、1500 米中长距离类，5000 米、10000 米、马拉松长距离类。

2. 样品聚类 (Q 型聚类)

例 11.2 为了更深入了解我国人口的文化程度状况, 现利用 1990 年全国人口普查数据对全国 30 个省、直辖市、自治区测试中的三个指标: (1) 大学以上文化程度的人口占全部人口的比例 (DXBZ); (2) 初中文化程度的人口占全部人口的比例 (CZBZ); 文盲半文盲人口占全部人口的比例 (WMBZ)、分别用来反映较高、中等、较低文化程度人口的状况, 原始数据见表 11-2, 数据已存放在 data11-02.sav 中。试对这 30 个省市进行聚类分析。(数据来源:《中国计划生育全书》P886)

在 SPSS 中进行类似本例作样品聚类的步骤如下:

1. 在数据编辑窗口中, 打开 data11-02.sav。

2. 按 Analyze→Classify→Hierarchical Cluster 顺序, 打开 Hierarchical Cluster Analysis 对话框, 见图 11-1。

表 11-2 30 个省市 3 项综合指标测试值

地区	序号	DXBZ	CZBZ	WMBZ
北京	1	9.30	30.55	8.70
天津	2	4.67	29.38	8.92
河北	3	0.96	24.69	15.21
山西	4	1.38	29.24	11.30
内蒙	5	1.48	25.47	15.39
辽宁	6	2.60	32.32	8.81
吉林	7	2.15	26.31	10.49
黑龙江	8	2.14	28.46	10.87
上海	9	6.53	31.59	11.04
江苏	10	1.47	26.43	17.23
浙江	11	1.17	23.74	17.46
安徽	12	0.88	19.97	24.43
福建	13	1.23	16.87	15.63
江西	14	0.99	18.84	16.22
山东	15	0.98	25.18	16.87
河南	16	0.85	26.55	16.15
湖北	17	1.57	23.16	15.79
湖南	18	1.14	22.57	12.10
广东	19	1.34	23.04	10.45
广西	20	0.79	19.14	10.61
海南	21	1.24	22.53	13.97
四川	22	0.96	21.65	16.24
贵州	23	0.78	14.65	24.27
云南	24	0.81	13.85	25.44
西藏	25	0.57	3.85	44.43
陕西	26	1.67	24.36	17.62
甘肃	27	1.10	16.85	17.93
青海	28	1.49	17.76	17.70
宁夏	29	1.61	20.27	22.06
新疆	30	1.85	20.66	12.75

在左侧的变量名源框中, 同时选择 DXBZ、CZBZ、WMBZ 变量, 单击右移箭头按钮, 将它们移入 Variable(s) 矩形框中。选择地区 (必须是字符型) 变量移入到 Label Cases By 框中。

由于本例要求做样品聚类, 故在 Cluster 栏中选择 Cases。

在 Display 栏中, 有两项选项, 必须选择一项, 本例中选择 Plots, 同上题一样, 主要考虑分类时用图示比较直观。

3. 单击 Plots 按钮, 打开 Hierarchical Cluster Analysis Plots 对话框, 见图 11-2。选择 Dendrogram, 要求输出树形图。

由于样品较多, 因此, 在 Icicle (冰柱图形) 多项选择中, 选择 None。不输出冰柱图。

单击 Continue 按钮, 返回 Hierarchical Cluster Analysis 对话框 (见图 11-1)。

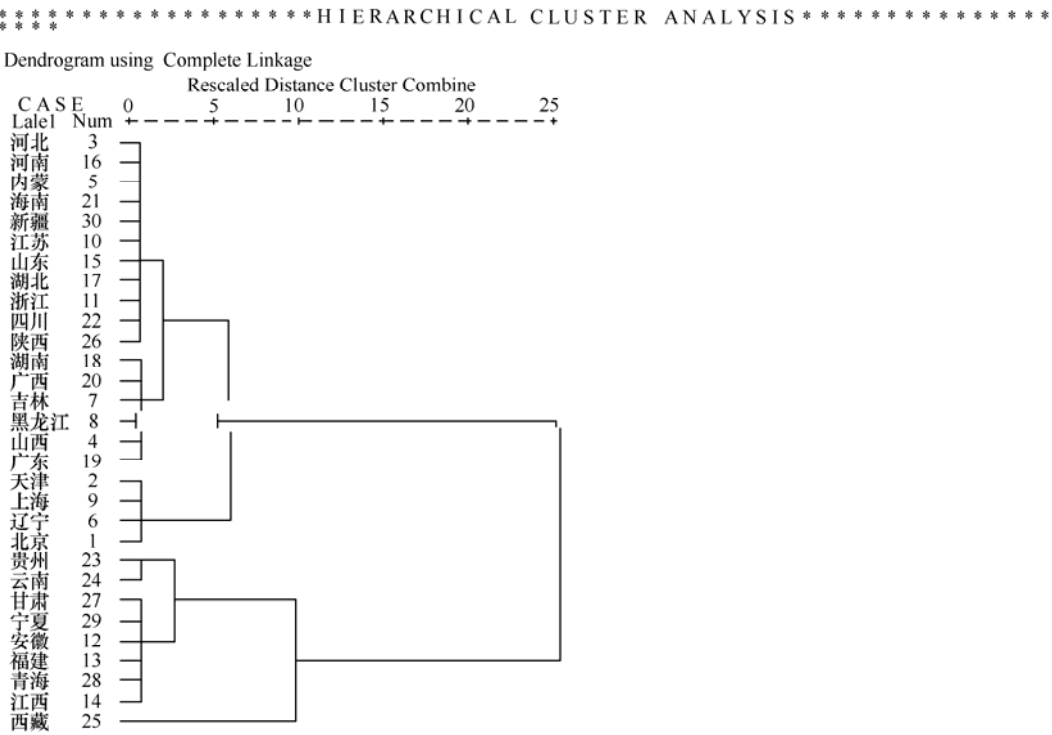
4. 单击 Method 按钮, 打开 Hierarchical Cluster Analysis: Method 对话框, 见图 11-3。在 Cluster Method (聚类方法) 中, 选择 Furthest neighbor (最长距离)。

在 Measure 中, 因本例的数据是尺度类型, 因此, 只能选择第一项 Interval。由于是进行样品聚类故选择 Cosine。

在 Transform Values 选项中，选择 Z scores。

其他采用系统默认值，单击 Continue 按钮，返回 Hierarchical Cluster Analysis 对话框（见图 11-1）。

5. 单击 OK 按钮执行，输出窗口中，出现聚类结果树状图，见图 11-9。



6. 结果。

在 2 左右将树状断开，可将 30 个省、市分成 4 类：

第一类：北京、辽宁、上海、天津，经济文化相对发达地区。

第二类：甘肃、宁夏、青海、安徽、浙江、四川、陕西、福建、江西。经济文化发展缓慢地区。

第三类：西藏，经济文化相对落后地区。

第四类：其他省、市、自治区。经济文化中等地区。

3. 数据资料为相关系数矩阵时的聚类分析

例 11.3 请将表 11-3 中的 10 个指标进行聚类分析。表中 10 个指标的实际含义分别为：X1—1000 米跑，X2—5 分钟往返跑，X3—爬竿，X4—举重，X5—跳远，X6—50 米跑，X7—引体向上，X8—立定跳远，X9—立卧撑，X10—砂袋掷远

表 11-3 10 指标的相关系数表

	1	2	3	4	5	6	7	8	9
1	1.00								
2	0.77	1.00							
3	0.30	0.73	1.00						
4	0.35	0.30	0.23	1.00					
5	0.35	0.27	0.13	0.21	1.00				
6	0.30	0.32	0.10	0.10	0.76	1.00			
7	0.10	0.16	0.10	0.39	0.21	0.10	1.00		
8	0.34	0.47	0.32	0.24	0.64	0.71	0.19	1.00	
9	0.35	0.49	0.34	0.21	0.66	0.73	0.36	0.60	1.00
10	0.30	0.27	0.19	0.70	0.20	0.17	0.70	0.22	0.20

对本类型的问题在 SPSS 中的解题步骤如下：

1. 在数据窗中直接输入矩阵数据

在数据窗中定义下列变量：

变量 Rowtype_，变量类型为短字符串型（变量长度≤8 个字符）。

在本变量名下，存放各观测量的统计学名称，其值必须使用规范的统计量名称。

常用规范的统计量名称有：

- ① MEAN 表示各变量均值。
- ② STDDEV（或 SD）表示各变量的标准差。
- ③ N（或 N_VECTOR）表示各变量的非缺失值的数目。
- ④ CORR 表示相关矩阵的相关系数。
- ⑤ COV 表示协方差阵的系数。
- ⑥ DEF 表示自由度。
- ⑦ MSE 表示误差的均方。

本例，在该变量下只要存放相关系数，因此，输入 CORR。

（1）变量 VARNAME_，变量类型为短字符串型（变量长度≤8 个字符），它的值为参与分析的（矩阵中涉及的）各变量的变量名。

本例中，输入需要分析的 10 个变量名，它们分别是 x1、x2、x3、x4、x5、x6、x7、x8、x9 和 x10。

（2）分析变量，变量类型为数字型，变量名分别是 x1、x2、x3、x4、x5、x6、x7、x8、x9 和 x10。

各变量名下输入的内容详见 data11-03.sav。

2. 按 Analyze→Classify→Hierarchical Cluster 顺序，打开 Hierarchical Cluster Analysis（分层聚类，也就是常说的系统聚类）对话框，见图 11-1。

在左侧的变量名框中，同时选择 x1、x2、x3、x4、x5、x6、x7、x8、x9 和 x10 变量，单击右移箭头按钮，将它们移入 Variable(s) 矩形框中。

由于本例要做指标（变量）聚类，因此，在 Cluster 栏中选择 Variables，要求进行 R 型指标聚类。

在 Display 栏中, 有两项选项, 必须选择一项, 本例中选择 Plots, 主要考虑分类时用图示比较直观。

3. 单击 Plots 按钮, 打开 Hierarchical Cluster Analysis: Plots 对话框, 见图 11-2。选择 Dendrogram, 要求输出树形图。

在 Icicle (冰柱图形) 多项选择中, 选择 None。不输出冰柱图。

单击 Continue 按钮, 返回 Hierarchical Cluster Analysis 对话框 (见图 11-1)。

4. 单击 OK 按钮执行, 则在输出窗口中, 得到如图 11-10 所示的聚类图。

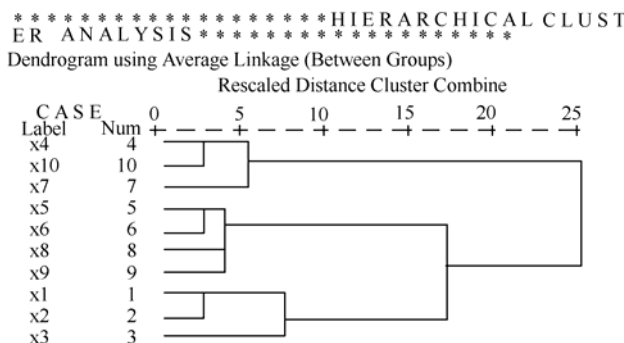


图 11-10 聚类图

5. 结果与分析。

在距离为 6 处将树状断开, 可将 10 个指标分为 3 类:

X4—举重、X10—砂袋掷远和 X7—引体向上为一类, 它为绝对力量类。

X5—跳远、X6—50 米跑、X8—立定跳远和 X9—立卧撑为一类, 它为速度爆发力类。

X1—1000 米跑、X2—5 分钟往返跑、X3—爬竿为一类, 它为耐力类。

11.1.4 典型指标的选择

当在一类中指标较多时, 一般我们会选取其中的一个作为典型指标, 这样, 一方面可以起到缩减同类指标的目的, 另一方面, 用选取的各类典型指标去做其他多元分析时, 就可避免多重共线性问题。

1. 选择典型指标的计算公式

$$\bar{R}_j^2 = \frac{\sum r^2}{m_j - 1}$$

式中 $\sum r^2$ 为第 j 类除计算典型指标的变量外的其他变量与计算典型指标的变量间的相关系数平方和, m_j 为第 j 类变量数。

2. 判定方法

根据计算公式 $\bar{R}_j^2 = \frac{\sum r^2}{m_j - 1}$ 计算得到类中指标的 \bar{R}^2 值最大者为该类典型指标。

3. 实例分析

例 11.4 在例 11.1 中，通过聚类分析得到，在长距离类中共有 5000 米、10000 米、马拉松长三个指标，试求其典型指标。

操作方法如下：

(1) 按 Analyze→Correlate→Bivariate 顺序单击菜单项，展开 Bivariate Correlate（相关分析）对话框。

在左侧变量名源中，选择@5000 米、@10000 米、马拉松变量，用右移箭头将其移入 Variables 框中。其他保持系统默认值。

单击 OK 按钮提交运行，得到表 11-4 相关矩阵表。

表 11-4 相关系数表

Correlations		@5000米	@10000米	马拉松
@5000米	Pearson Correlation	1	.975 ^{**}	.932 [*]
	Sig. (2-tailed)		.000	.000
	N	55	55	55
@10000米	Pearson Correlation	.975 ^{**}	1	.943 [*]
	Sig. (2-tailed)	.000		.000
	N	55	55	55
马拉松	Pearson Correlation	.932 ^{**}	.943 ^{**}	1
	Sig. (2-tailed)	.000	.000	
	N	55	55	55

**. Correlation is significant at the 0.01 level (2-tailed).

(2) 从表 11-4 中读取相关系数，计算各相关指数：

$$\bar{R}_{5000\text{米}}^2 = (0.975^2 + 0.932^2) / 2 = 0.909445$$

$$\bar{R}_{10000\text{米}}^2 = (0.975^2 + 0.943^2) / 2 = 0.919748$$

$$\bar{R}_{\text{马拉松}}^2 = (0.932^2 + 0.943^2) / 2 = 0.879278$$

(3) 结论

由于，在三个相关指数中， $\bar{R}_{10000\text{米}}^2 = 0.919748$ 最大，所以，10000 米成绩指标是这三个指标中的典型指标。

11.1.5 动态聚类分析

在大样本情况下，系统聚类分析需要占用大量的计算机内存，计算工作量大和需要较多计算时间，因此，从节省成本，提高效率的目的出发，需要一种能在大样本条件下替代系统聚类分析的快速而又对计算机内存要求相对较低的方法，动态聚类分析正是在这种需求下应运而生的一种新的聚类方法，由于它的计算速度快，因此也称它为快速聚类分析。

11.1.5.1 基本思想

动态聚类的基本思想为：它对数据先进行初始分类或选择一批凝聚点（类中心点），让变量（或样品）按某种准则向凝聚点凝聚，然后逐步调整，直至分类比较合理或迭代稳定得到最终分类为止。

动态聚类速度的快慢取决于使用者的经验，如果使用者经验丰富，则在指定初始的类中心点时就比较合理，这样可以减少运算时大量的迭代次数，从而可以缩短计算时间。

动态聚类有很多方法，这里主要介绍与 SPSS 中 K-Mean 程序相一致的由麦奎因提出

的 K -均值法。其基本步骤如下：

1. 将所有样品分成 k 个初始类，然后用这 k 个类的重心（均值）作为初始凝聚点，或选择 k 个样品作为初始凝聚点（类中心点）。
2. 样品间远近程度一般用欧氏距离测度，计算所有样品数据点到 k 个类中心点的欧氏距离。将每个样品按距 k 个类中心点距离最近的准则，把所有样品归入离类中心点最近的那个类中，形成一个新的 k 类，完成一次迭代过程。
3. 重新计算 k 个类的中心点，以其作为新的类中心点。
4. 重复 2、3 步骤，直至达到终止迭代的条件或指定的迭代次数为止。

11.1.5.2 实例分析

例 11.5 从 1996 年第 1 期《世界经济统计研究》中，摘录得到 1992 年以下 14 个国家的出生时的预期寿命（岁）、成人识字率（%）、调整后人均 GDP，见表 11-5，数据存放在 data11-04.sav 中。试对 14 个国家进行快速聚类分析。

表 11-5 1992 年 14 个国家在三项指标上的观测数据资料

编号	国家名称	预期寿命	成人识字率	人均 GDP
1	美国	76	99	5374
2	日本	79.5	99	5359
3	瑞士	78	99	5372
4	阿根廷	72.1	95.9	5242
5	阿联酋	73.8	77.7	5370
6	保加利亚	71.2	93	4250
7	古巴	75.3	94.9	3412
8	巴拉圭	70	91.2	3390
9	格鲁吉亚	72.8	99	2300
10	南非	62.9	80.6	3799
11	中国	68.5	79.3	1950
12	罗马尼亚	69.9	96.9	2840
13	希腊	77.6	93.8	5233
14	哥伦比亚	69.3	90.3	5158

在 SPSS 中进行类似本例作样品聚类的步骤如下：

1. 在数据编辑窗口中，打开 data11-04.sav。
2. 按 Analyze→Classify→K-Mean Cluster 顺序，打开 K-Mean Cluster Analysis 对话框，见图 11-11。

在左侧的变量名源框中，同时选择 *预期寿命*、*识字率*、*人均 GDP* 变量，单击右移箭

头按钮，将它们移入 Variable(s)矩形框中。选择国家名称（必须是字符型）变量移入到 Label Cases By 框中。

在 Number of Clusters 框中输入 2（这是系统默认值）。

3. 单击 Option 按钮，展开 K-Mean Cluster Analysis Option 对话框，见图 11-12。

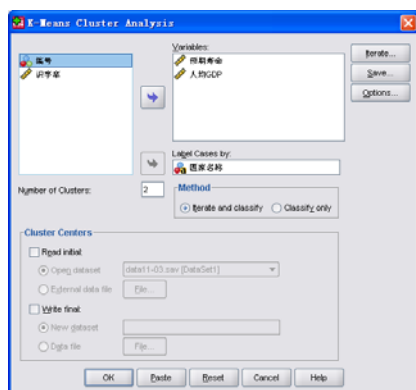


图 11-11 K-Mean Cluster Analysis 对话框

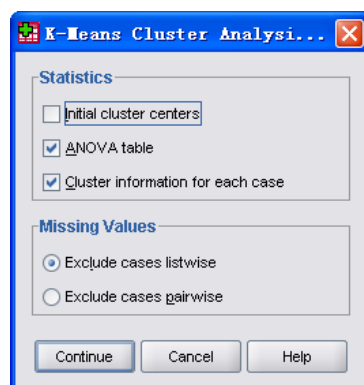


图 11-12 K-Mean Cluster Analysis Option 对话框

在 Statistics 选项中，选择 ANOVA table 和 Cluster information for each case 选项，要求输出方差分析表和对每个样品的聚类信息。

单击 Continue 返回 K-Mean Cluster Analysis 对话框。

4. 单击 OK 按钮运行，则在输出窗口中得到 6 张表。

5. 结果与分析。

现先分析输出窗口中的第 5 张方差分析表，见表 11-6。

表 11-6 方差分析表

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
预期寿命	78.583	1	15.129	12	5.194	.042
识字率	33.930	1	59.419	12	.571	.464
人均GDP	1.692E7	1	297716.125	12	56.833	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

从表 11-6 中可见，聚成的两类在预期寿命、识字率和人均 GDP 指标的均值间无差异的原假设下，出现目前值或更极端值的概率分别为 0.042、0.464 和 0.000，故变量预期寿命和人均 GDP 在分类过程中均在统计学上有显著性意义（ $P=0.042$ 和 0.00 都小于 0.05 ），而识字率变量无统计学上的显著性意义（ $P=0.464$ 大于 0.05 ），因此，有必要在剔除识字率变量后重新做聚类分析。

除在第 2 步中，只选择预期寿命、人均 GDP 变量到 Variable(s)矩形框中，其他保持

不变，重复上述步骤后，可在输出窗口中得到 6 张表，分别见表 11-7 至表 11-22。

表 11-7 显示了迭代过程的基本情况，它表明迭代到第三次时，类中心点已没有太大变化达到收敛。

表 11-8 显示各国分成两类的情况以及各样品离其类中心点的距离。由该表可见，美国、日本、瑞士、阿根廷、阿联酋、保加利亚、希腊和哥伦比亚 8 国被分成一类，它们是高发展水平国家，古巴、巴拉圭、格鲁吉亚、南非、中国和罗马尼亚 6 国被分成一类，它们是中等发展水平国家。

表 11-7 迭代过程记录

Iteration History ^a			
Iteration	Change in Cluster...		
	1	2	
1	356.343	828.405	
2	152.339	170.106	
3	.000	.000	

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 3424.008.

表 11-8 分类情况

Cluster Membership			
Case...	国家名称	Cluster	Distance
1	美国	1	204.004
2	日本	1	189.061
3	瑞士	1	202.027
4	阿根廷	1	72.046
5	阿联酋	1	200.002
6	保加利亚	1	918.007
7	古巴	2	463.531
8	巴拉圭	2	441.500
9	格鲁吉亚	2	648.506
10	南非	2	850.529
11	中国	2	998.501
12	罗马尼亚	2	108.500
13	希腊	1	63.067
14	哥伦比亚	1	13.154

表 11-9 显示了最终类中心点，第一类的类中心点为：预期寿命为 74.69、人均 GDP 为 5170，第二类的类中心点为：预期寿命为 69.90、人均 GDP 为 2948.5。

表 11-10 显示了最终两类中心点间的欧氏距离为 2.222×10^3 。

表 11-9 最终类中心点

	Cluster	
	1	2
预期寿命	74.69	69.90
人均GDP	5170.00	2948.50

表 11-10 最终两类中心点间的欧氏距离

Distances between Final Cluster Centers		
Cluster	1	2
1		2.222E3
2	2.222E3	

表 11-11 给出了分类变量的方差分析表，从表中可见用来聚类的两个变量预期寿命和人均 GDP 在分两类过程中，均有统计学上的显著性意义 ($P=0.042$ 和 0.000 均小于 0.05)，表明用这两个变量将样品分成两类的快速聚类过程是成功的，聚类效果有统计学意义。

表 11-12 给出了各类中样品数量信息，聚到第一类的共有 8 个国家，聚到第二类的共

表 11-11 方差分析表

ANOVA						
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
预期寿命	78.583	1	15.129	12	5.194	.042
人均GDP	1.692E7	1	297716.125	12	56.833	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

表 11-12 在各类中的样品数量

Number of Cases in each Cluster		
Cluster	1	8.000
	2	6.000
Valid		14.000
Missing		.000

有 6 个国家，有效数为 14，没有缺失值情况。

11.1.6 判别分析

11.1.6.1 判别分析概述

1. 何谓判别分析

判别分析是根据已知类别的事物性质来建立判别函数，然后据此对未知类别的新事物进行判断，将其归入到已知类别中的一种统计分析方法。

由于判别问题在各行各业、包括在我们的日常生活中都非常常见，例如，你在河中抓到一条未见过的鱼，要判别它属于何种鱼，又如，在运动员选材研究中，通常需要判定一名运动员最适合的运动项目，以便在训练中能起到事半功倍的功效等，可以说在各个研究领域中都要用到判别分析。因而，从应用的频繁程度而言，它完全可以同回归分析相媲美。

判别分析和前面介绍的聚类分析都是用来分类的一种统计方法。所不同的是，在进行聚类分析前，样品的分类情况可以是未知的，因而要利用未知类别的样品间的相似性或接近性，根据一定的准则来将它们进行分类。而判别分析是在已知研究对象分类及已取得各类的一批已知样品观测数据的基础上，根据某些准则建立起判别式（函数），然后用判别式对未知类型的样品进行判别分类。因此，两种分析方法的用途间有很大的区别。但两者间又是相互依赖，互为补充的。聚类分析与判别分析往往联合起来应用，因为，在很多情况下，总体的分类是不清楚的，因而在做判别分析前，首先要通过聚类分析来进行分类，然后再用判别分析建立判别式对新样品进行判别。

2. 判别分析的内容

判别分析的内容十分丰富。按判别的组数来分：有两组判别分析和多组判别分析；按区分不同总体的数学模型来分：有线性判别和非线性判别；按判别时所处理的变量方法的不同来分：有逐步判别和序贯判别等。判别准则也丰富多彩：有马氏距离最小准则、Fisher 准则、平均损失最小准则、最小平方准则、最大似然准则、最大概率准则等。因而，按判别准则的不同又产生了多种判别方法。

3. 判别分析需要满足的一些条件

在判别分析中，通常把预报因子看成是随机向量，每一个类看成是一个总体。如同其他统计方法一样，判别分析也有一些条件要求：

(1) 预测变量原则上是连续型或有序分类变量，因变量是分类变量。如果预测变量中有名义多分类变量时，需将其转换成哑元后再一起进行判别分析。

(2) 在 Bayes 判别分析中，要求样本来自一个多元正态总体，但在 Fisher 判别分析中，这一条件不需要。

(3) 各组的协方差矩阵相等，类似于方差分析中的方差齐性。

(4) 预测变量之间没有显著的相关, 也即无共线性, 预测变量的均值与方差不相关。在不同的类中, 两个预测变量间的相关性应是一样的。

(5) 为使建立的判别函数比较稳定, 样本量应为建立判别函数的自变量的 10~20 倍。

相对而言, 判别分析在违反这些适用条件时显得非常稳健, 它们对结果的影响不是很大。事实上, 在很多实际问题中, 要求样本来自一个多元正态总体, 几乎是做不到的。因此, 保守一些的话, 可以用 Fisher 判别法即可。

11.1.6.2 Fisher 判别法

Fisher (费歇) 判别法是 1936 年由 Fisher 首先提出的, 其优势在于对总体的分布和方差等都没有特殊要求, 因而得到广泛应用。在理论上可以证明, Fisher 判别法同典型相关分析是等价的, 故也称它为典型判别或典则判别。

1. 基本思想

Fisher 判别法同主成分分析有关, 更等价于典型相关分析, 其基本思想是首先提取出与各组有最大可能多重相关的变量的线性组合, 即第一典型变量, 然后再提取第二典型变量, 一般提取到两至三个典型变量后即可, 再用典型变量计算出个类别在低维空间中的重心坐标, 给出判别式用以计算各样品的坐标值, 最后用各观测点离各类重心距离的远近来做出样品所属类别的判断。

2. 判别函数的建立

设有 k 个总体 G_1, \dots, G_k , 从每个总体 G_i 中抽取 n_i 个样品, $i = 1, 2, \dots, k$, 令总的样本量为 $n = \sum_i n_i$ 。每个样品测得 p 个指标, 则 $x_{\alpha}^{(i)} = (x_{\alpha 1}^{(i)}, \dots, x_{\alpha p}^{(i)})$ 为第 i 个总体的第 α 个样品的观测向量。则实测结果的记录形式见表 11-13 所示。

表 11-13 实测结果记录表

G_1 总体					\dots					G_k 总体				
变量 \ 样品	x_1	x_2	\dots	x_p	\dots					变量 \ 样品	x_1	x_2	\dots	x_p
$x_1^{(1)}$	$x_{11}^{(1)}$	$x_{12}^{(1)}$	\dots	$x_{1p}^{(1)}$	\dots					$x_1^{(i)}$	$x_{i1}^{(i)}$	$x_{i2}^{(i)}$	\dots	$x_{ip}^{(i)}$
$x_2^{(1)}$	$x_{21}^{(1)}$	$x_{22}^{(1)}$	\dots	$x_{2p}^{(1)}$	\dots					$x_2^{(i)}$	$x_{i2}^{(i)}$	$x_{i2}^{(i)}$	\dots	$x_{ip}^{(i)}$
\vdots	\vdots	\vdots	\vdots	\vdots	\dots					\vdots	\vdots	\vdots	\vdots	\vdots
$x_{n_1}^{(1)}$	$x_{n_1 1}^{(1)}$	$x_{n_1 2}^{(1)}$	\dots	$x_{n_1 p}^{(1)}$	\dots					$x_{n_i}^{(i)}$	$x_{n_i 1}^{(i)}$	$x_{n_i 2}^{(i)}$	\dots	$x_{n_i p}^{(i)}$
均值	$\bar{x}_1^{(1)}$	$\bar{x}_2^{(1)}$	\dots	$\bar{x}_p^{(1)}$	\dots					均值	$\bar{x}_1^{(i)}$	$\bar{x}_2^{(i)}$	\dots	$\bar{x}_p^{(i)}$

设所要建立的判别函数为

$$y(x) = c_1 x_1 + \dots + c_p x_p = c'x$$

其中, $c = (c_1, c_2, \dots, c_p)$, $x = (x_1, x_2, \dots, x_p)$ 。

记 $\bar{x}^{(i)} = (\bar{x}_1^{(i)}, \dots, \bar{x}_p^{(i)})'$, $i=1, 2, \dots, k$, 和 $s^{(i)} = \sum_{\alpha=1}^{n_i} (X_{\alpha}^{(i)} - \bar{X}^{(i)})(X_{\alpha}^{(i)} - \bar{X}^{(i)})'$ 为总体 G_i 内 x 的样本均值向量和样本协差阵, 由此可得 $y(x)$ 在 G_i 上的样本均值和样本方差为

$$\bar{y}^{(i)} = c' \bar{x}^{(i)}, \quad \sigma_i^2 = c' s^{(i)} c$$

记 \bar{x} 为总的均值向量, 则 $\bar{y} = c' \bar{x}$ 。

为了使判别函数能很好地区别来自不同总体的样品, 必须使来自不同总体的组间离差相差越大越好; 对于分属不同总体的各组的组内离差越小越好。

也就是要求

$$\lambda = \frac{\sum_{i=1}^k n_i (\bar{y}^{(i)} - \bar{y})^2}{\sum_{i=1}^k q_i \sigma_i^2}$$

越大越好。式中, q_i 是人为设定的大于 0 的加权系数, 它可以取为先验概率。一般都假定各总体的先验概率相等。

为使后面的推导运算过程变得简单些, 一般可取 $q_i = n_i - 1$, 则上式可简化为

$$\lambda = \frac{c' A c}{c' E c}$$

式中, $E = \sum_{i=1}^k q_i s^{(i)}$ 为组内离差阵, $A = \sum_{i=1}^k n_i (\bar{x}^{(i)} - \bar{x})(\bar{x}^{(i)} - \bar{x})'$ 为总体之间的协差阵。

因此, 在多总体下, Fisher 的准则就是要选取能使上式达到最大的系数向量 c 。根据极限存在的必要条件, 令 $\frac{\partial \lambda}{\partial c} = 0$ 可得, $A c = \lambda E c$ 。这表明, λ 及 c 恰好就是 A 、 E 矩阵的广义特征根及其对应的特征向量。

由于 E 和 A 通常都能满足正定的条件, 因此, 有不超过 $m = \min(k-1, p)$ 个非零特征根的存在, 记为 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m > 0$, 因而可以构造 m 个判别函数

$$y_l(x) = c^{(l)'} x \quad l=1, 2, \dots, m$$

为对每个判别函数的判别能力的大小作出评判, 定义

$$p_l = \frac{\lambda_l}{\sum_{i=1}^m \lambda_i} \quad (l=1, 2, \dots, m)$$

即用非零特征根与非零特征根的和之比值作为衡量每个判别函数判别能力的指标。 p 值越大, 对应判别函数的判别能力越强。

m_0 个判别函数 y_1, y_2, \dots, y_{m_0} 的判别能力定义为

$$sp_{m_0} = \sum_{l=1}^{m_0} p_l$$

一般由上式求得值大于人为的指定值（如大于 0.85）时，此时得到 m_0 个判别函数就足够了。

3. 分类

(1) 当 $m_0=1$ 即只取一个判别函数时，有两种可选的分类方法：

① 不加权法

如果 $y(x) - \bar{y}^{(i)} = \min_{1 \leq j \leq k} |y(x) - \bar{y}^{(j)}|$ ，则判 $x \in G_i$ 。

② 加权法

将 $\bar{y}^{(1)}, \bar{y}^{(2)}, \dots, \bar{y}^{(k)}$ 由小到大排列成 $\bar{y}_{(1)} \leq \bar{y}_{(2)} \leq \dots \leq \bar{y}_{(k)}$ ，对其标准差也相应重排为 $\sigma_{(i)}$ 。

定义 G_{j_i} 与 $G_{j_{i+1}}$ 间的分界点为

$$d_{i,i+1} = \frac{\sigma_{(i+1)}\bar{y}_{(i)} + \sigma_{(i)}\bar{y}_{(i+1)}}{\sigma_{(i+1)} + \sigma_{(i)}} \quad i = 1, 2, \dots, k-1$$

若 x 能使 $d_{i-1,i} \leq y(x) \leq d_{i,i+1}$ ，则判 $x \in G_{j_i}$ 。

(2) 当 $m_0 > 1$ 时，同样也有两种可选的分类方法：

① 不加权法

记 $\bar{y}_l^{(i)} = c^{(l)}, \bar{x}^{(i)} \quad l = 1, 2, \dots, m_0; i = 1, 2, \dots, k$ ，对待判样品 $x = (x_1, x_2, \dots, x_p)$ 计算

$$y_l(x) = c^{(l)}, x, \quad D_i^2 = \sum_{l=1}^{m_0} [y_l(x) - \bar{y}_l^{(i)}]^2 \quad i = 1, 2, \dots, k$$

若 $D_r^2 = \min_{1 \leq i \leq k} D_i^2$ ，则判 $x \in G_r$ 。

② 加权法

由于各判别函数判别能力的不同，因此，定义 $D_i^2 = \sum_{l=1}^{m_0} [y_l(x) - \bar{y}_l^{(i)}]^2 \lambda_l$ 作为统计量，

其中 λ_l 就是上面提到的 A 、 E 矩阵的广义特征根。

若 $D_r^2 = \min_{1 \leq i \leq k} D_i^2$ ，则判 $x \in G_r$ 。

4. 实例分析

例 11.6 为建立男子少年游泳员的选材模式，观测了仰泳男子少年运动员和蛙泳男子少年运动员各 10 名的力量（上肢拉力/体重）、肩幅和两足外展角度三项指标的数据资料，见表 11-14，数据存放在 data11-05.sav 中。试对其求两个项目的 Fisher 判别式。并对力量

(上肢拉力/体重)、肩幅和两足外展角度分别为 47.06、5 和 176 的新样品进行判别。

表 11-14 仰泳和蛙泳男子少年运动员三项素质、形态指标的测试结果

组别 1 (仰泳)			组别 2 (蛙泳)		
力量	肩幅	两足外展角度	力量	肩幅	两足外展角度
50.74	6.0	170.0	50.25	9.3	190.0
47.06	3.0	180.0	44.76	13.0	186.0
57.45	3.2	174.0	37.63	8.4	180.0
50.25	5.4	160.0	47.06	11.6	192.0
49.4	4.0	172.0	43.8	7.9	178.0
48.99	4.5	168.0	50.0	6.0	176.0
55.23	6.5	170.0	39.2	3.5	182.0
46.63	3.1	182.0	41.0	7.0	176.0
51.28	6.2	176.0	46.5	6.8	184.0
50.19	5.4	168.0	42.9	5.6	180.0

在 SPSS 中的具体解题步骤如下：

1. 打开分析的数据文件

在 SPSS 数据编辑窗口中，打开数据文件 data11-05.sav。

2. 定义分类因变量和自变量

按 Analyze→Classify→Discriminant 顺序单击菜单项，展开 Discriminant Analysis (判别分析) 对话框，见图 11-13。

在 Discriminant Analysis 对话框中左面的矩形框中选择 *组别* 变量，并将其送入 Grouping Variable 框中。此时矩形框下面的 Define Range 按钮加亮，单击该按钮，打开定义分类变量范围的对话框，如图 11-14 所示。

在 Minimum 框中输入该分类变量的最小值 1，在 Maximum 框中输入该分类变量的最大值 2。表明要分两类。

单击 Continue 按钮返回 Discriminant Analysis 对话框。

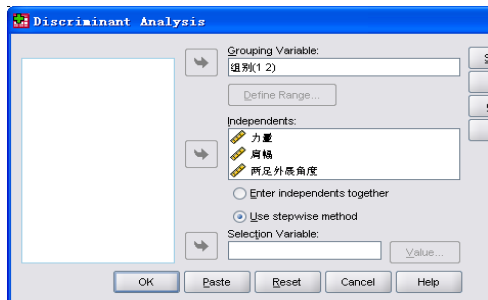


图 11-13 Discriminant Analysis 对话框

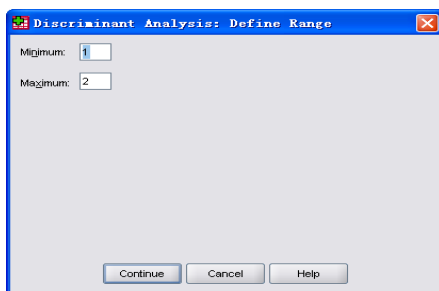


图 11-14 定义分类变量范围的对话框

在 Discriminant Analysis 对话框左面的变量表中选择 *力量*、*肩幅* 和 *两足外展角度* 变量，

将它们送到 Independents 矩形框中，作为参与判别分析的自变量。

3. 选择分析方法

在对话框中自变量矩形框下面有两个选择项，(1) Enter independent together 选项，(2) Use stepwise method 选项，可供我们用来选择判别分析方法。方法(2)是逐步判别法，类似回归分析中逐步回归法，将在另一节中介绍。本例选择方法(1)，即使用所有自变量进行判别分析，建立全模型。

4. 指定输出的统计量

单击 Statistics 按钮，展开 Discriminant Analysis: Statistics 对话框，如图 11-15 所示。

在 Descriptives 栏中选择 Univariate ANOVA 选项，要求进行假设检验，输出单变量的方差分析结果。检验的无效假设是：各类中同一自变量均值都相等。

值得一提的是，Box's M 选项是用来对各类的协方差阵相等的假设进行检验的一种方法。

由于本例我们要用 Fisher 判别法建立判别式，所以它对总体分布和方差等都没有具体要求，因而可以不选择 Box's M 选项。

在 Function coefficients 栏中选择 Unstandardized 复选项，输出未经标准化的 Fisher 判别系数。

同样，需要提醒的是，如果选择 Fisher's 选项，给出的是 Bayes 判别函数的系数，而不是 Fisher 判别函数的系数。由于按判别函数值最大的一组进行归类的思想是 Fisher 提出来的，所以，为了纪念 Fisher 对判别分析作出的贡献，SPSS 的程序编制者们才会如此命名。

单击 Continue 按钮返回 Discriminant Analysis 对话框。

5. 指定分类参数和判别结果

在对话框中单击 Classify 按钮，展开 Discriminant Analysis: Classification 对话框，如图 11-16 所示。

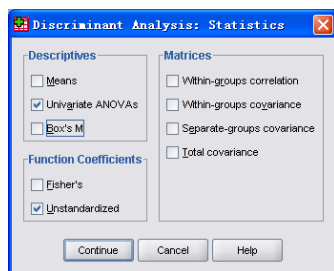


图 11-15 Discriminant Analysis: Statistics 对话框

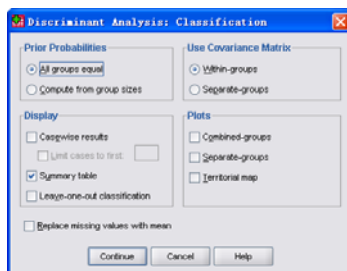


图 11-16 Discriminant Analysis: Classification 对话框

在 Display 栏中选择 Summary table 要求输出分类小结，给出正确分类观测量数，即

原始类和根据判别函数计算的预测类相同的观测量数、错分观测量数和错分率。

其他保持系统默认选项。单击 Continue 按钮返回 Discriminant Analysis 对话框。

6. 输出结果及其讨论

在 Discriminant Analysis 对话框中, 单击 OK 按钮提交执行, 则在输出窗口中得到输出结果, 分别见表 11-15 至表 11-26。

表 11-15 分析样品处理摘要

Analysis Case Processing Summary		
Unweighted Cases		
Valid		20 100.0
Excluded	Missing or out-of-range group codes	0 .0
	At least one missing discriminating variable	0 .0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0 .0
	Total	0 .0
Total		20 100.0

表 11-16 分组统计

组别		Valid N (listwise)	
		Unweighted	Weighted
仰泳	力量	10	10.000
	肩幅	10	10.000
	两足外展角度	10	10.000
蛙泳	力量	10	10.000
	肩幅	10	10.000
	两足外展角度	10	10.000
Total	力量	20	20.000
	肩幅	20	20.000
	两足外展角度	20	20.000

表 11-15 显示共有 20 个样品 (没有缺失值) 作为判别基础数据参与到判别分析中。

表 11-16 显示分组统计情况, 共有两组, 一组为仰泳组, 一组为蛙泳组, 测试指标为力量、肩幅和两足外展角度, 每组样品数都为 10。

表 11-17 自变量在各组中均值相等检验

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
力量	.564	13.939	1	18	.002
肩幅	.637	10.276	1	18	.005
两足外展角度	.544	15.059	1	18	.001

表 11-18 特征根

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2.595 ^a	100.0	100.0	.850

a. First 1 canonical discriminant functions were used in the analysis.

表 11-17 显示了力量、肩幅和两足外展角度三个自变量在各组中均值相等的检验情况, 它相当于做了单因素方差分析。从表中可见, 这三个自变量在仰泳和蛙泳两个组的均值间的差异, 都有统计上的极显著性意义 ($P=0.002$ 、 0.005 、 0.001 均小于 0.01)。说明用这三个自变量建立的判别式是有足够的判别能力的。

表 11-18 显示, 在本例中只需一个判别函数, 其特征根为 $\lambda_1 = 2.595$, 该函数的判别能力为 100%, 判别函数与组别间的关联程度为 0.850 (典型相关系数)。

表 11-19 Wilks Λ 值

Wilks' Lambda				
Test of ...	Wilks' Lambda	Chi-square	df	Sig.
1	.278	21.112	3	.000

表 11-20 标准化典型判别函数系数

Standardized Canonical Discriminant Function Coefficients	
	Function
力量	-.769
肩幅	.600
两足外展角度	.526

表 11-19 显示了 Wilk 的 Λ 值, 它和表 11-18 中的特征根有如下关系

$$Wilks' \Lambda = \frac{1}{1 + \lambda_1} = \frac{1}{1 + 2.595} = 0.278$$

表中的卡方按下式计算， $df = p(k - 1)$ ， p 为指标数， k 为分组数，本例分别为 3 和 2。

$$\chi^2 = \left(n - 1 - \frac{p + k}{2} \right) \ln \Lambda \sim \chi^2(p(k - 1))$$

卡方检验结果表明，拒绝仰泳和蛙泳两组多元总体均值相等的原假设（ $P=0.000<0.05$ ）。表明判别函数有很强的判别能力，也就是说所建立的判别函数在统计学上是有显著性意义的。本检验一般用在 Bayes 判别分析的检验中，具体要求参见下一节检验部分的内容。

表 11-20 给出了标准化的 Fisher 判别函数的系数，由此可得标准化 Fisher 判别函数为

$$F = -0.769x_1 + 0.600x_2 + 0.526x_3$$

式中 x_1 为标准化的力量数据、 x_2 为标准化的肩幅数据和 x_3 为标准化的两足外展角度数据。

要用该判别函数做新样品的判别分析时，首先要将新样品数据做标准化处理，标准化处理的计算公式为 $z = \frac{x - \bar{x}}{s}$ ，由于要涉及原始数据的转换，所以，实际分析中一般不用该判别函数。

表 11-21 结构矩阵表

Structure Matrix	
	Function
	1
两足外展角度	.568
力量	-.546
肩幅	.469

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

表 11-22 典型判别函数系数表

Canonical Discriminant Function Coefficients	
	Function
	1
力量	-.200
肩幅	.270
两足外展角度	.088
(Constant)	-7.748

Unstandardized coefficients

表 11-21 给出的是结构矩阵表，它是变量和判别函数的组内相关矩阵，相关系数越大表明该变量对判别函数影响越大。

表 11-22 显示未标准化的 Fisher 判别函数的系数。由此可得 Fisher 判别式为

$$F = -7.748 - 0.200x_1 + 0.270x_2 + 0.088x_3$$

式中 x_1 为力量的原始数据、 x_2 为肩幅的原始数据和 x_3 为两足外展角度的原始数据。

表 11-23 显示了各类重心在空间中的坐标位置。此处为 1 维，如仰泳的重心位置为（-1.528），这样，只要在计算前面计算出各观测的具体坐标位置后，再计算出它们分别离各重心的距离，就可以得知它们的分类了。

表 11-23 各类重心

Functions at Group Centroids	
组别	Function
	1
仰泳	-1.528
蛙泳	1.528

Unstandardized canonical discriminant functions evaluated at group means

表 11-24 分类处理摘要

Classification Processing Summary	
Processed	20
Excluded	Missing or out-of-range group codes 0
	At least one missing discriminating variable 0
Used in Output	20

表 11-24 显示的是分类处理信息。在分类处理中共有 20 个样品，全部样品都用来做判别函数，中间没有缺失样品的情况发生。

表 11-25 用于分类的先验概率

Prior Probabilities for Groups			
组别	Prior	Cases Used in Analysis	
		Unweighted	Weighted
仰泳	.500	10	10.000
蛙泳	.500	10	10.000
Total	1.000	20	20.000

表 11-26 分类结果

Classification Results ^a				
Original	Count	Predicted Group Membership		Total
		组别 仰泳	蛙泳	
		10	0	10
		1	9	10
	%	100.0	0	100.0
		10.0	90.0	100.0

a. 95.0% of original grouped cases correctly classified.

表 11-25 显示各类的先验概率都为 0.50，使用于判别分析的未加权样品数两组都为 10，加权样品数也都为 10。

表 11-26 给出了分类结果，从表中可见，用原有样品回代到求得的判别函数中，仰泳组没有 1 个错判，而蛙泳组有 1 人被判到仰泳组，表明判别函数对仰泳组的判别能力为 100%，对蛙泳组的判别能力为 90%，总的回代正确率为 95%。

将力量（上肢拉力/体重）、肩幅和两足外展角度分别为 47.06、5 和 176 的新样品的数据代入到 Fisher 判别函数 $F = -7.748 - 0.200x_1 + 0.270x_2 + 0.088x_3$ 得到 $F = -0.322$ ，由于它离仰泳重心的距离近（ $1.206 < 1.85$ ，或 $-0.322 < 0$ ），所以，判新样品属于第一类，即该运动员适合练仰泳。

11.1.6.3 多总体 Bayes 判别法

1. Bayes（贝叶斯）判别法的基本思想

当总体有一定先验概率时，先验概率对判别的影响较大。Bayes 判别将所有类看作是空间中互斥的子域，而每个观测都是空间中的一个点。它是建立在对已知研究对象的先验概率的基础上，利用 Bayes 公式按照一定准则构建一个判别函数，然后用它计算待判样品属于各总体的条件概率，通过比较它在各个总体中的条件概率的大小，将待判样品判归为来自概率最大的总体的一种判别方法。同 Fisher 判别法相比，Bayes 判别更适合于多总体判别。但正如上面已提到过的，Bayes 判别适用于多元正态总体，并要求多元正态总体的协方差阵相等。

2. 多元正态总体的 Bayes 判别法

(1) 判别函数和判别准则

设有 k 个总体 G_1, \dots, G_k ，它们分别服从多元正态分布 $N(\mu_1, \Sigma_1)$ ， $N(\mu_2, \Sigma_2)$ ， \dots ， $N(\mu_k, \Sigma_k)$ ，它们的先验概率分别为 q_1, q_2, \dots, q_k ， $\sum_{i=1}^k q_i = 1$ 。当先验概率未知时，可用样品频率代替， $q_g = \frac{n_g}{n}$ ，其中 n_g 为建立判别函数的已知分类数据中，来自第 g 总体的

样品的数量, 而 $n = \sum_{i=1}^g n_i$, 即来自所有总体的样品总数, 当然也可以令先验概率相等, 都等于 $\frac{1}{k}$, 此时, 可以认为先验概率不起作用。

对于多元正态总体, 其概率密度函数为

$$f_g(x) = (2\pi)^{-p/2} |\Sigma^{(g)}|^{-1/2} \cdot \exp \left\{ -\frac{1}{2} (x - \mu^{(g)})' \Sigma^{(g)-1} (x - \mu^{(g)}) \right\}$$

其中, $\mu^{(g)}$ 和 $\Sigma^{(g)}$ 分别是第 g 总体的 p 维均值向量和 p 阶协差阵。

于是对于随机抽取的一个向量 x (即随机观测到一个样品的情况下), 它属于第 g 个总体的概率为

$$P(g/x) = \frac{q_g f_g(x)}{\sum_{i=1}^k q_i f_i(x)}$$

当我们将 $f_g(x)$ 代入上式后可知, 不论 g 怎么变化, 分式中的分母它总是常量, 于是按使 $P(h/x)$ 为最大值的分类法 (称最大概率法), 则等价于

$$q_g f_g(x) \xrightarrow{g} \max$$

则判 x 来自第 g 总体。

对上式取对数并去掉与 g 无关的项, 则得

$$\begin{aligned} Z(g/x) &= \ln q_g - \frac{1}{2} \ln |\Sigma^{(g)}| - \frac{1}{2} (x - \mu^{(g)})' \Sigma^{(g)-1} (x - \mu^{(g)}) \\ &= \ln q_g - \frac{1}{2} \ln |\Sigma^{(g)}| - \frac{1}{2} x' \Sigma^{(g)-1} x - \frac{1}{2} \mu^{(g)'} \Sigma^{(g)-1} \mu^{(g)} + x' \Sigma^{(g)-1} \mu^{(g)} \end{aligned}$$

因而, 问题就转化为

$$Z(g/x) \xrightarrow{g} \max$$

为进一步简化计算的工作量, 在满足 $\Sigma^{(1)} = \Sigma^{(2)} = \dots = \Sigma^{(k)} = \Sigma$ (协方差阵相等) 时, 在 $Z(g/x)$ 中的 $\frac{1}{2} \ln |\Sigma^{(g)}|$ 和 $\frac{1}{2} x' \Sigma^{(g)-1} x$ 两项都与 g 无关, 因此, 在求最大值时可以将这两项去掉。由此可得到下列的判别函数与判别准则

$$\begin{cases} y(g/x) = \ln q_g - \frac{1}{2} \mu^{(g)'} \Sigma^{(g)-1} \mu^{(g)} + x' \Sigma^{(g)-1} \mu^{(g)} \\ y(g/x) \xrightarrow{g} \max \end{cases}$$

(2) 后验概率的计算

计算后验概率的公式为

$$P(g/x) = \frac{\exp\{y(g/x)\}}{\sum_{i=1}^k \exp\{y(i/x)\}}$$

3. 统计检验

设从每个总体 G_i 中抽取 n_i 个样品, $i=1,2,\dots,k$, 令总的样本量为 $n = \sum_i^k n_i$ 。每个样品测得 p 个指标。实测结果的记录形式见表 11-13 所示。

令总的离差阵为 T

$$T = \sum_{i=1}^k \sum_{t=1}^{n_i} (x_t^{(i)} - \bar{x})(x_t^{(i)} - \bar{x})'$$

k 个总体的组间离差阵为 A

$$A = \sum_{i=1}^k n_i (\bar{x}^{(i)} - \bar{x})(\bar{x}^{(i)} - \bar{x})'$$

总体 r 的组内离差阵为 E_r

$$E_r = \sum_{t=1}^{n_r} (x_t^{(r)} - \bar{x}^{(r)})(x_t^{(r)} - \bar{x}^{(r)})'$$

$$E = E_1 + E_2 + \dots + E_k$$

记

$$A = \frac{|E|}{|T|} = \frac{|E|}{|A+E|}$$

称它为 Wilks 统计量, 因为它有 p 个变量组成, 所以也记为 $A_{(p)}$ 。

(1) 对 k 个多元正态总体均值的检验

$$H_0: \mu^{(1)} = \mu^{(2)} = \dots = \mu^{(k)}$$

① 卡方检验

在大样本情况下, 在原假设为真时, 统计量

$$\chi^2 = \left(n - 1 - \frac{p+k}{2} \right) \ln \Lambda \sim \chi^2(p(k-1))$$

当 $P(H_0) < 0.05$ 时, 拒绝均值都相等的原假设, 而认为至少有两个均值间是不等的。

② Rao 检验

在原假设为真时, 统计量

$$F \approx \frac{Ds - 2\lambda}{p(k-1)} \cdot \frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \sim F(p(k-1), Ds - 2\lambda)$$

式中,

$$D = n - 1 - \frac{1}{2}(p + k), \quad \lambda = \frac{p(k-1) - 2}{4}$$

$$s = \begin{cases} \frac{p^2(k-2)^2 - 4}{p^2 + (k-1)^2 - 5}, & \text{当 } p^2 + (k-1)^2 - 5 \neq 0 \\ 1, & \text{当 } p^2 + (k-1)^2 - 5 = 0 \end{cases}$$

当 $P(H_0) < 0.05$ 时, 拒绝均值都相等的原假设, 而认为至少有两个均值间是不等的。

③ 适用于非正态分布的 V 统计量

它由 Mardia 于 1971 年提出, 在原假设为真时

$$F \approx \frac{v_2}{v_1} \cdot \frac{v^*}{1 - v^*} \sim F(df_1, df_2)$$

式中,

$$V = \sum_{r=1}^k n_r (\bar{\mathbf{x}}^{(r)} - \bar{\mathbf{x}})' \mathbf{T}^{-1} (\bar{\mathbf{x}}^{(r)} - \bar{\mathbf{x}})$$

$$a = \min(p, k-1), \quad v_1 = p(k-1), \quad v_2 = a(n-1) - p(k-1)$$

$$v^* = \frac{V}{a}, \quad 0 < v^* < 1$$

$$df_1 = \delta v_1 = \delta p(k-1), \quad df_2 = \delta v_2 = \delta \{a(n-1) - p(k-1)\}$$

其中 δ 称为非正态性的调整因子, 其计算公式为

$$R_i = \sum_{r=1}^k \frac{1}{n_r} - \frac{k^2}{n}$$

$$b = \frac{1}{n} \sum_{r=1}^k \sum_{t=1}^{n_r} \left\{ (\bar{\mathbf{x}}^{(r)} - \bar{\mathbf{x}}) \left(\frac{\mathbf{T}}{n} \right)^{-1} (\bar{\mathbf{x}}^{(r)} - \bar{\mathbf{x}}) \right\}^2$$

$$C_x = \frac{n-1}{(k-1)(n-3)(n-k)} \{n(n+1)R_1 - 2(k-1)(n-k)\}$$

$$C_y = \frac{n-1}{p(n-3)(n-p-1)} \{(n+1)b - p(n-1)(p+2)\}$$

$$\Gamma_y = E(C_y)$$

$$c = \frac{(n-3)C_x \Gamma_y}{2n(n-1)}, \quad \delta^{-1} = 1 + \frac{c\{a(n-1) + 2\}}{a(n-1) - 2c}$$

当 \mathbf{x} 为正态分布且协方差相等时, 有 $\Gamma_y = 0$, 即 $\delta = 1$ 。在实际使用时, 用 C_y 取代 Γ_y 。

(2) 对每个变量的检验

当要检验变量 x_i 在 k 个总体间的条件均值相等时, 可用统计量

$$F = \frac{1 - \Lambda_{i,(p-1)}}{\Lambda_{i,(p-1)}} \cdot \frac{n - p - k + 1}{k - 1} \sim F(k - 1, n - p - k + 1)$$

其中, $\Lambda_{i,(p-1)} = \frac{t^{(ii)}}{e^{(ii)}}$, 而 $t^{(ii)}$ 及 $e^{(ii)}$ 为 T^{-1} 及 E^{-1} 阵的第 i 对角线元素。

当 $P(H_0) < 0.05$ 时, 拒绝变量 x_i 在 k 个总体间的条件均值都相等的原假设, 而认为至少在两个总体上变量 x_i 的均值间是不等的。

(3) 总体间的差异性检验

所要检验的原假设为 H_0 : 总体 G_i 与 G_j 之间无差异, 在原假设为真时, 统计量

$$F = \frac{(n - k - p + 1)}{(n - k)(n_i + n_j)p} D_{ij}^2 \sim F(p, n - p - k + 1) \quad i \neq j = 1, 2, \dots, k$$

其中 D_{ij}^2 为从 G_i 与 G_j 中随机抽取的样本的马氏距离。

4. 实例分析

例 11.7 从胃癌患者、萎缩性胃炎患者和非胃病者中各随机抽取 5 人, 测试其铜兰蛋白、兰色反应、吡啶乙酸和中性硫化物指标的值, 测得的数据资料见表 11-27, 数据已存放在 data11-06.sav 中, 试对其用 Bayes 判别法进行判别分析。

表 11-27 3 组不同患者在 4 个测试指标上的测试值

		铜兰蛋白	兰色反应	吡啶乙酸	中性硫化物
胃癌患者	1	228	134	0.20	0.11
	2	245	134	0.10	0.40
	3	200	167	0.12	0.27
	4	170	150	0.07	0.08
	5	100	167	0.20	0.14
萎缩性胃炎患者	6	225	125	0.07	0.14
	7	130	100	0.06	0.12
	8	150	117	0.07	0.06
	9	120	133	0.10	0.25
	10	160	100	0.05	0.10
非胃病者	11	185	115	0.05	0.19
	12	170	125	0.06	0.04
	13	165	142	0.05	0.03
	14	135	108	0.02	0.12
	15	100	117	0.07	0.02

在 SPSS 中的解题步骤如下:

(1) 打开分析的数据文件

在 SPSS 数据编辑窗口中, 打开数据文件 data11-06.sav。

(2) 对铜兰蛋白、蓝色反应、吲哚乙酸和中性硫化物 4 个自变量做正态性检验
用第 2 章数据资料探索性分析中介绍的方法，可得铜兰蛋白、蓝色反应、吲哚乙酸和中性硫化物 4 个自变量的正态性检验结果，见表 11-28。

表 11-28 数据资料的正态性检验

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
铜兰蛋白	.127	15	.200 [*]	.959	15	.675
蓝色反应	.138	15	.200 [*]	.940	15	.380
吲哚乙酸	.287	15	.002	.807	15	.004
中性硫化物	.228	15	.034	.888	15	.063

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

从表中可见，4 个自变量在各组的观察的显著性水平均大于 0.05，故现有证据不足以拒绝这 4 个变量服从正态分布的原假设。

(3) 定义分类因变量和自变量

按 Analyze→Classify→Discriminant 顺序单击菜单项，展开 Discriminant Analysis（判别分析）对话框，见图 11-13。

在 Discriminant Analysis 对话框中左面的矩形框中选择 *胃病类型* 变量，并将其送入 Grouping Variable 框中。此时矩形框下面的 Define Range 按钮加亮，单击该按钮，打开定义分类变量范围的对话框，如图 11-14 所示。

在 Minimum 框中输入该分类变量的最小值 1，在 Muximum 框中输入该分类变量的最大值 3。表明要分三类。

单击 Continue 按钮返回 Discriminant Analysis 对话框。

(4) 选择分析方法

在对话框中自变量矩形框下面选择(1) Enter independent together 选项，即使用所有自变量进行判别分析，建立全模型。

(5) 指定输出的统计量

单击 Statistics 按钮，展开 Discriminant Analysis: Statistics 对话框，如图 11-15 所示。

在 Descriptives 栏中选择 Univariate ANOVA 选项，要求进行假设检验，输出单变量的方差分析结果。检验的无效假设是：各类中同一自变量均值都相等。

同时选择 Box's M 选项，要求对各类的协方差阵相等的假设进行检验的一种方法。

在 Function coefficients 栏中选择 Fisher's 选项，要求给出 Bayes 判别函数的系数。单击 Continue 按钮返回 Discriminant Analysis 对话框。

(6) 指定分类参数和判别结果

在对话框中单击 Classify 按钮，展开 Discriminant Analysis: Classification 对话框，如图 11-17 所示。

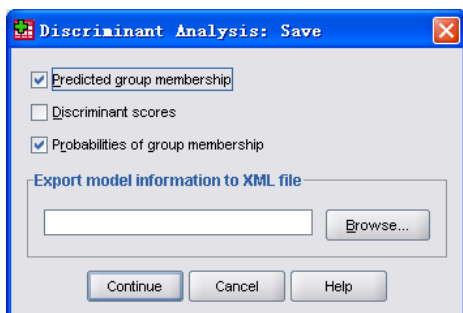


图 11-17 Discriminant Analysis: Save 对话框

在 Display 栏中选择 Summary table 选项，要求输出分类小结，给出正确分类观测量数，即原始类和根据判别函数计算的预测类相同的观测量数、错分观测量数和错分率。

其他保持系统默认选项。单击 Continue 按钮返回 Discriminant Analysis 对话框。

(7) 在工作的数据文件中存取分类结果及其概率

单击 Save 按钮，展开 Discriminant Analysis: Save 对话框，见图 11-17。选择 Predicted group membership 和 Probabilities of group membership 选项，要求在工作数据文件中存取预测的分类情况和分到各类的概率值。

(8) 输出结果及其讨论

在 Discriminant Analysis 对话框中，单击 OK 按钮提交执行，则在输出窗口中得到输出结果，分别见表 11-29 至表 11-42。

表 11-29 分析样品处理摘要

Analysis Case Processing Summary			
Unweighted Cases		N	Percent
Valid		15	100.0
	Missing or out-of-range group codes	0	.0
	At least one missing discriminating variable	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
	Total	0	.0
Total		15	100.0

表 11-30 分组统计

Group Statistics			
		Valid N (listwise)	
		Unweighted	Weighted
胃癌	铜兰蛋白	5	5,000
	蓝色反应	5	5,000
	吡啶乙酸	5	5,000
	中性硫化物	5	5,000
萎缩性胃炎	铜兰蛋白	5	5,000
	蓝色反应	5	5,000
	吡啶乙酸	5	5,000
	中性硫化物	5	5,000
没有胃病	铜兰蛋白	5	5,000
	蓝色反应	5	5,000
	吡啶乙酸	5	5,000
	中性硫化物	5	5,000
Total	铜兰蛋白	15	15,000
	蓝色反应	15	15,000
	吡啶乙酸	15	15,000
	中性硫化物	15	15,000

表 11-29 显示共有 15 个样品（没有缺失值）作为判别基础数据参与到判别分析中。

表 11-30 显示分组统计情况，共有三组，一组为胃癌患者，一组为萎缩性胃炎患者，另一组为非胃病者，测试指标为铜兰蛋白、蓝色反应、吡啶乙酸和中性硫化物，每组样品数都为 5。

表 11-31 显示了铜兰蛋白、蓝色反应、吡啶乙酸和中性硫化物四个自变量在各组中均

表 11-31 自变量在各组中均值相等检验

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
铜兰蛋白	.857	1.003	2	12	.396
蓝色反应	.426	8.074	2	12	.006
吡啶乙酸	.442	7.564	2	12	.007
中性硫化物	.762	1.879	2	12	.195

表 11-32 对数决定系数

Log Determinants		
胃癌类型	Rank	Log Determinant
胃癌	4	2.522
萎缩性胃炎	4	-6.184
没有胃病	4	-3.106
Pooled within-groups	4	1.498

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

值相等的检验情况。从表中可见，蓝色反应、吡啶乙酸自变量在胃癌、萎缩性胃炎和非胃病者三个类别的均值间的差异有统计上的显著性意义 ($P=0.006$ 、 0.007)，另两个变量没有统计上的显著性意义 ($P=0.396$ 、 0.195)。说明用这四个自变量建立的判别式从纯理论的角度而言并不是最佳的。但在变量（测试指标）不多的情况下，更应该首先考虑判别式的判别效果，这是做判别分析的目的。

表 11-32 和表 11-33 是选择 Box's M 选项的结果，主要用来检验协方差阵是否是齐性的。因此，可以把重点放在表 11-33 上。从表 11-33 可见， $P=0.357>0.05$ ，表明没有足够的理由拒绝协方差阵齐性的假设。

由此可见，用来做判别分析的 4 个自变量是基本满足变量的分布正态和方差齐性的假设的。

表 11-33 协方差阵相等检验

Test Results		
Box's M		45.046
F	Approx.	1.089
	df1	20
	df2	516.896
	Sig.	.357

Tests null hypothesis of equal population covariance matrices.

表 11-34 特征根

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	3.044 ^a	93.6	93.6	.868
2	.207 ^a	6.4	100.0	.414

a. First 2 canonical discriminant functions were used in the analysis.

表 11-34 分别给出了两个判别函数的特征根（也称特征值）、占方差的百分比、占方差的累计百分比、典型相关系数。其中第一个判别函数解释了所有变异的 93.6%，剩下的 6.4 的变异则由第二个判别函数来解释。

由于本例中分类数为 3，因此，判别函数为 $3-1=2$ 个，判别函数的效应量可以用典型相关系数的平方来评定。显而易见，第一个判别函数的效应量 ($0.868^2=0.753424$) 要远高于第二个判别函数 ($0.414^2=0.171396$)。

表 11-35 Wilks' Λ 值

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.205	16.649	8	.034
2	.828	1.978	3	.577

表 11-36 标准化典型判别函数系数

	Function	
	1	2
铜兰蛋白	.453	-.175
蓝色反应	.596	-.811
吡啶乙酸	.662	.600
中性硫化物	.299	.608

表 11-35 显示了 Wilk 的 Λ 值，它和表 11-34 中的特征根有如下关系：

$$\text{判别函数 1 中 Wilks' Lambda} = \frac{1}{(1 + \lambda_1)(1 + \lambda_2)} = \frac{1}{(1 + 3.044)(1 + 0.207)} = 0.205$$

$$\text{判别函数 2 中 Wilks' Lambda} = \frac{1}{(1 + \lambda_2)} = \frac{1}{(1 + 0.207)} = 0.828$$

$$df_{1-2} = p(k-1) = 4 * 2 = 8, \quad df_2 = (p-1)(k-2) = 3 * 1 = 3$$

卡方值的计算可以参见对表 11-19 中的说明。

卡方检验结果表明，两个判别函数的平均数在三个组别间的差异有统计学上的显著性意义 ($P=0.034<0.05$)。而在排除第一个判别函数后，第二个函数在三个组别间的差异没有统计学上的显著性意义 ($P=0.577>0.05$)。

当然，我们还可以使用偏 $\eta^2 = 1 - \Lambda^{1/3}$ 来计算这个分析的整体效应量。

表 11-36 给出了标准化的 Bayes 判别函数的系数，由此可得标准化 Bayes 判别函数为

$$F_1 = -0.453x_1 + 0.596x_2 + 0.662x_3 + 0.299x_4$$

$$F_2 = -0.175x_1 - 0.811x_2 + 0.600x_3 + 0.608x_4$$

式中， x_1 、 x_2 、 x_3 和 x_4 分别表铜兰蛋白、蓝色反应、吲哚乙酸和中性硫化物的标准化处理后的数据。

该表显示为最大程度区分各类每个变量被赋予多大的权重。因此，自变量前面的系数也是该自变量对判别函数所起贡献的权重大小的标志。

要用该判别函数做新样品的判别分析时，首先要把新样品数据做标准化处理，所以，实际分析中一般不用该判别函数。

表 11-37 结构矩阵

	Function	
	1	2
吲哚乙酸	.638 [*]	.327
铜兰蛋白	.234 [*]	.057
蓝色反应	.643	-.645 [*]
中性硫化物	.295	.478 [*]

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

表 11-38 各类重心

	Function	
	1	2
胃癌类型		
胃癌	2.199	-.049
萎缩性胃炎	-.936	.522
没有胃病	-1.263	-.472

Unstandardized canonical discriminant functions evaluated at group means

表 11-37 显示的是结构矩阵，它是变量和判别函数的组内相关矩阵，因此，这些数据是每个自变量与判别函数的相关系数。

表 11-38 列出了各类重心在二维空间中的坐标位置。如胃癌患者重心为 (2.199, -0.049)。这样，只要计算得到各观测具体坐标位置后，在计算出它们分别离各类重心的距离，就可以对它们进行分类了。

表 11-39 给出了分类处理摘要，在分类处理中共有 15 个样品，全部样品都用来做判别函数，中间没有缺失样品的情况发生。

表 11-39 分类处理摘要

Classification Processing Summary		
Processed		15
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		15

表 11-40 各类的先验概率

Prior Probabilities for Groups				
胃病类型	Prior	Cases Used in Analysis		
		Unweighted	Weighted	
胃癌	.333	5	5.000	
萎缩性胃炎	.333	5	5.000	
没有胃病	.333	5	5.000	
Total	1.000	15	15.000	

表 11-41 Bayes 分类函数系数

	胃病类型		
	胃癌	萎缩性胃炎	没有胃病
铜兰蛋白	.164	.130	.130
蓝色反应	.753	.595	.637
吡啶乙酸	77.825	31.661	9.990
中性硫化物	7.280	1.247	-5.920
(Constant)	-79.212	-46.721	-49.598

Fisher's linear discriminant functions

表 11-40 显示各类的先验概率都为 0.333，使用于判别分析的未加权样品数两组都为 5，加权样品数也都为 5。

表 11-41 给出了 Bayes 分类函数的系数。由此可得未标准化处理的 Bayes 判别函数为：

$$F_1 = -79.212 + 0.164x_1 + 0.753x_2 + 77.825x_3 + 7.280x_4$$

$$F_2 = -46.721 + 0.130x_1 + 0.595x_2 + 31.661x_3 + 1.247x_4$$

$$F_3 = -49.598 + 0.130x_1 + 0.637x_2 + 9.990x_3 - 5.920x_4$$

式中， F_1 为胃癌的判别函数， F_2 为萎缩性胃炎， F_3 为没有胃病的判别函数， x_1 、 x_2 、 x_3 和 x_4 分别表铜兰蛋白、蓝色反应、吡啶乙酸和中性硫化物指标。

利用这些判别函数就可以计算待判样品在各类上的得分，得分最高的一类就是该样品应归入的类别。

表 11-42 实际分类结果统计

Classification Results ^a					
		胃病类型	Predicted Group Membership		
			胃癌	萎缩性胃炎	没有胃病
Original	Count	胃癌	4	0	1
		萎缩性胃炎	0	4	1
		没有胃病	0	1	4
	%	胃癌	80.0	.0	20.0
		萎缩性胃炎	.0	80.0	20.0
		没有胃病	.0	20.0	80.0

a. 80.0% of original grouped cases correctly classified.

从表 11-42 中可见，5 个胃癌患者中正确分类率为 80%，有一个即 20%被错分到没有胃病类中，同样萎缩性胃炎患者类中 80%的被正确分类，有一个即 20%被错分到没有胃病类中，而在没有胃病类中正确分类率也是 80%，只有一个即 20%被错分到萎缩性胃炎类中。总的正确分类率为 80%。

在工作的数据文件中，新增四列变量，分别为 Dis_1、Dis1_1、Dis2_1 和 Dis3_1，在这四个变量中分别存放分类结果、分为第一类的概率、分为第二类的概率和分为第三类的概率。可见，在样品被归类到三类中概率最大的那个类中。

11.1.6.4 逐步判别法

无论是 Fisher 判别法，还是 Bayes 判别法，都是用给定的全部变量来建立判别式的，在判别式中，每个变量所起的作用实际上是不一样的，同回归分析一样，进入判别式的变量也应是重要的变量，或者说它应是在统计学上是有意义的变量。因此，必须对进入

判别式的变量要进行筛选。

1. 逐步判别法的基本思想

它同逐步回归法的基本思想类似，都是采用“有进有出”的算法，也就是在建立判别式时，不是一次将全部变量均纳入到判别式中，而是每步引入一个变量，每引入一个“最重要”的变量进入判别式时，同时考虑引入判别式的所有变量对判别式的影响，如果某个变量已不显著了，不管它是新引进的变量，还是判别式中的原有变量，都要将它从判别式中剔除，直到判别式中都是起作用的变量，而剩下来的变量也没有重要的变量可引入到判别式为止。

2. 引入和剔除变量所用的检验统计量

对总体的要求同 Bayes 判别法，现从 k 个总体中分别抽取 n_1, n_2, \dots, n_k 个样品，由此可得 $n = \sum_{i=1}^k n_i$ ，对每个样品测得 p 个指标，则 $x_{\alpha}^{(i)} = (x_{\alpha 1}^{(i)}, \dots, x_{\alpha p}^{(i)})$ 为第 i 个总体的第 α 个样品的观测向量。实测结果的记录形式见表 11-13 所示。

所要做的原假设为： $H_0: \mu^{(1)} = \mu^{(2)} = \dots = \mu^{(k)}$

检验用的似然比的统计量为

$$A_p = \frac{|E|}{|T|} = \frac{|E|}{|A+E|} \sim \Lambda_p(n-k, k-1)$$

显然， $0 < \Lambda_p < 1$ ，而 $|E|$ 和 $|A|$ 分别反映了同一总体样本间的差异和 k 个总体所有样本间的差异。所以， Λ_p 越小表明相同总体间的差异越小。当 $P(H_0) < 0.05$ 时，拒绝 H_0 。

在大样本的情况下，上面检验统计量也常用以下的近似公式来替代：

(1) Bartlett 近似式

$$-\left\{n - \frac{1}{2}(p-k) - 1\right\} \ln \Lambda \sim \chi^2(p(k-1))$$

(2) Rao 近似式

$$\frac{\{n - (p-1) - k\}}{k-1} \cdot \left(\frac{\Lambda_{p-1}}{\Lambda_p} - 1 \right) \sim F(k-1, n-p-k+1)$$

在引入变量时，检验统计量为

$$F_{lr} = \frac{1 - A_r}{A_r} \cdot \frac{n-l-k}{k-1} \sim F(k-1, n-l-k)$$

式中， l 为计算到第 l 步， $A_r = \frac{e_{rr}^l}{t_{rr}^l}$ ，而 $t^{(ii)}$ 及 $e^{(ii)}$ 为 T^{-1} 及 E^{-1} 阵的第 i 对角线元素。

如果 $F_{lr} > F_{\alpha}(k-1, n-l-k)$ ，则表明变量 x_r 判别能力显著，将判别能力显著的变量

中 F_{1r} 值最大的变量引入。在引入（或删除）变量后，都需要对相应的矩阵 E 和 T 做一次消去变换。

在剔除变量时，在已选入判别式的所有变量中，找出最大 A_r 的一个变量进行检验。检验用的统计量为

$$F_{2r} = \frac{1 - A_r}{A_r} \cdot \frac{n - L - 1 - k}{k - 1} \sim F(k - 1, n - L - k + 1)$$

式中， L 为进行到第 l 步时，判别式中含有的自变量数。

若 $F_{2r} < F_{\alpha}(k - 1, n - L - k + 1)$ ，则认为 x_r 的判别能力不显著，将其从判别式中剔除。

重复这两个步骤，直到判别式中都是起作用的变量，而剩下来的变量也没有重要的变量可引入到判别式为止。

3. 实例分析

例 11.8 仍以例 11.7 的资料为例，试对其用逐步判别法进行判别分析。

(1) 打开分析的数据文件

在 SPSS 数据编辑窗口中，打开数据文件 data11-06.sav。

(2) 定义分类因变量和自变量

按 Analyze→Classify→Discriminant 顺序单击菜单项，展开 Discriminant Analysis（判别分析）对话框，见图 11-13。

在 Discriminant Analysis 对话框中左面的矩形框中选择 **肝病类型** 变量，并将其送入 Grouping Variable 框中。此时矩形框下面的 Define Range 按钮加亮，单击该按钮，打开定义分类变量范围的对话框，如图 11-14 所示。

在 Minimum 框中输入该分类变量的最小值 1，在 Maximum 框中输入该分类变量的最大值 3。表明要分三类。

单击 Continue 按钮返回 Discriminant Analysis 对话框。

(3) 选择分析方法

在对话框中自变量矩形框下面选择 Use stepwise method 选项，即使用逐步判别法进行判别分析，建立判别模型。

选择本选项后，Method 按钮被激活，单击 Method 按钮，展开 Discriminant Analysis: Stepwise Method 对话框，如图 11-18 所示。

在本对话框中可选择所用的方法和变量进出的准则。本例选用系统默认选项。

单击 Continue 按钮返回 Discriminant Analysis 对话框。

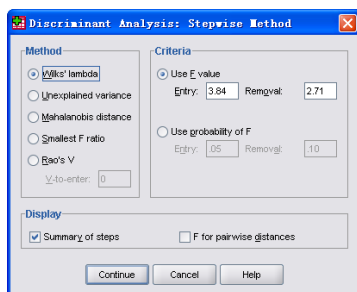


图 11-18 Discriminant Analysis: Stepwise Method 对话框

(4) 指定输出的统计量

单击 Statistics 按钮, 展开 Discriminant Analysis: Statistics 对话框, 如图 11-15 所示。

在 Descriptives 栏中选择 Univariate ANOVA 选项, 要求进行假设检验, 输出单变量的方差分析结果。检验的无效假设是: 各类中同一自变量均值都相等。

同时选择 Box's M 选项, 要求对各类的协方差阵相等的假设进行检验的一种方法。

在 Function coefficients 栏中选择 Fisher's 选项, 要求给出 Bayes 判别函数的系数。

单击 Continue 按钮返回 Discriminant Analysis 对话框。

(5) 指定分类参数和判别结果

在对话框中单击 Classify 按钮, 展开 Discriminant Analysis: Classification 对话框, 如图 11-16 所示。

在 Display 栏中选择 Summary table 选项, 要求输出分类小结, 给出正确分类观测量数, 即原始类和根据判别函数计算的预测类相同的观测量数、错分观测量数和错分率。

其他保持系统默认选项。单击 Continue 按钮返回 Discriminant Analysis 对话框。

(6) 输出结果及其讨论

单击 OK 按钮提交执行, 则在输出窗口中得到输出结果, 其中前 2 张表同表 11-29 和表 11-30, 解释略。

其他表分别见表 11-43 至表 11-50。

表 11-43 变量的选入和剔除汇总

Variables Entered/Removed ^{a, b, c, d}									
Step	Entered	Wilks' Lambda				Exact F			
		Statistic	df1	df2	df3	Statistic	df1	df2	Sig.
1	蓝色反应	.426	1	2	12.000	8.074	2	12.000	.006

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 8.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

d. F level, tolerance, or VIN insufficient for further computation.

从表 11-43 可见, 每一步中最小 Wilks' Lambda 对应的变量被选入。最多步数为 8。变量的选入和从判别方程中剔除变量的标准为: 进入判别方程的变量的最小 F 值为 3.84, 从判别方程中剔除变量的最大 F 值为 2.71。

从逐步判别分析运行的记录中可知, 它只运行了一步, 也即只选入了一个蓝色反应变量进入判别方程, 同时 F 检验结果显著 ($P=0.006<0.05$), 说明只有蓝色反应变量对正确分类结果起作用。

表 11-44 已在判别分析中的变量情况

Variables in the Analysis		
Step	Tolerance	F to Remove
1 蓝色反应	1.000	8.074

表 11-45 未在判别分析中的变量情况

Variables Not in the Analysis				
Step	Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0				
铜兰蛋白	1.000	1.000	1.003	.857
蓝色反应	1.000	1.000	8.074	.426
吡啶乙酸	1.000	1.000	7.564	.442
中性硫化物	1.000	1.000	1.879	.762
1				
铜兰蛋白	.944	.944	1.190	.350
吡啶乙酸	.935	.935	2.451	.295
中性硫化物	.998	.998	1.228	.349

表 11-44 至表 11-46 也是逐步判别分析的运行记录，它补充说明了逐步判别分析的情况。表 11-44 分别给出了进入判别方程的变量蓝色反应的容忍度（1.000）和 F 值（8.074），表 11-45 则详细记录了在各步中没有进入判别方程的变量情况、容忍度值、最小容忍度值、F 值和 Wilks' Lambda 值。而表 11-46 是对逐步判别分析的小结，同表 11-43。

表 11-46 Wilks' Lambda

Wilks' Lambda									
Step	Number of Variables	Lambda	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	1	.426	1	2	12	8.074	2	12.000	.006

表 11-47 特征根

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.346 ^a	100.0	100.0	.757

a. First 1 canonical discriminant functions were used in the analysis.

表 11-48 Wilks' Lambda

Wilks' Lambda				
Test of ...	Wilks' Lambda	Chi-square	df	Sig.
1	.426	10.231	2	.006

表 11-49 标准化典型判别函数系数

Standardized Canonical Discriminant Function Coefficients	
	Function
	1
蓝色反应	1.000

表 11-50 结构矩阵

Structure Matrix	
	Function 1
蓝色反应	1.000
吡啶乙酸 ^a	.255
铜兰蛋白 ^a	-.237
中性硫化物 ^a	-.048

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.
a. This variable not used in the analysis.

表 11-51 各类重心

Functions at Group Centroids	
	Function
	1
胃痛类型	
胃癌	1.446
萎缩性胃炎	-.939
没有胃病	-.507

Unstandardized canonical discriminant functions evaluated at group means

表 11-52 分类处理摘要

Classification Processing Summary	
Processed	15
Excluded	
Missing or out-of-range group codes	0
At least one missing discriminating variable	0
Used in Output	15

表 11-53 各类的先验概率

Prior Probabilities for Groups			
胃病类型	Prior	Cases Used in Analysis	
		Unweighted	Weighted
胃癌	.333	5	5.000
萎缩性胃炎	.333	5	5.000
没有胃病	.333	5	5.000
Total	1.000	15	15.000

表 11-54 Bayes 判别函数系数

Classification Function Coefficients			
	胃病类型		
	胃癌	萎缩性胃炎	没有胃病
蓝色反应	.682	.522	.551
(Constant)	-52.423	-31.105	-34.538

Fisher's linear discriminant functions

对表 11-47 至表 11-55 的解释参见对表 11-34 至表 11-42 的说明。

由此可得未标准化处理的 Bayes 判别函数为

$$F_1 = -52.423 + 0.682x_2$$

$$F_2 = -31.105 + 0.522x_2$$

$$F_3 = -34.538 + 0.551x_2$$

式中, F_1 为胃癌的判别函数, F_2 为萎缩性胃炎, F_3 为没有胃病的判别函数, x_2 表蓝色反应指标。

表 11-55 分类结果

Classification Results ^a					
		Predicted Group Membership			Total
		胃癌	萎缩性胃炎	没有胃病	
Original	Count				
	胃癌	3	0	2	5
	萎缩性胃炎	0	3	2	5
	%				
	胃癌	60.0	.0	40.0	100.0
	萎缩性胃炎	.0	60.0	40.0	100.0
	没有胃病	20.0	60.0	20.0	100.0

a. 46.7% of original grouped cases correctly classified.

从表 11-55 可见, 各类的分类正确率分别为 60%、60% 和 20%, 全部分类的正确率为 46.7%。

通过例 11.7 和例 11.8 的对比不难看出, 当不考虑变量对判别函数的显著性时, 可以得到较高的分类的正确率, 而在判别方程中只选入有显著作用的变量时, 判别的正确率则在明显下降。

一般而言, 当判别的正确率没有太多的下降时, 使用逐步判别分析多少可限制进入判别方程的自变量的数量, 这对实际操作无疑是有作用的, 但判别分析追求的是判别的正确率, 因此, 在遇到本例中的特殊情况, 也就是鱼和熊掌不能得全时, 就不一定要追求统计上的完美, 应选择有更高分类正确率的做法。

第 12 章 主成分分析、因子分析与对应分析

在讨论的变量较多时，一些变量之间难免存在相关性，因此，人们希望能够找出它们的少数“代表”，也就是内在因子，或潜在的内在结构（概念变量）来对它们进行描述，这样可以降低讨论问题时的复杂性，以便于描述、理解和分析众多的变量。主成分分析（Principal Component Analysis）和因子分析正是在这种想法下应运而生的两种对变量进行降维处理的统计方法。

主成分分析和因子分析在概念上是有区别的。主成分分析只是想用少量的综合指标来获取与大量变量等价的信息，而因子分析通过对变量内在结构的研究去了解变量间的关系。由于在数据处理的形式上，两种方法极为相似，因此，实际工作者和一些理论家常把两者不加区别。

使用主成分分析和因子分析进行处理多变量之间的关系时，有两个前提条件，第一个是变量之间必须相关，第二个条件是要有较大的样本含量。样本含量的大小同测试指标数（即变量的数量）有关，具体要求参见相关与回归分析。同样，它们都需要数据资料服从正态分布。

主成分分析和因子分析都有分析变量间关系的 R 型分析和分析样品间关系的 Q 型分析。而将 R 型因子分析和 Q 型因子分析结合起来进行分析的统计分析方法就是对应分析。

本章主要对它们的基本概念做一些解释，重点放在如何用 SPSS 的相应程序进行统计处理，及对 SPSS 的输出结果进行解释上。

12.1 主成分分析

12.1.1 主成分分析及其基本思想

1. 主成分分析的基本概念

在实际研究中，经常会遇到多指标的问题，由于不同的指标间多少存在一定的相关性，这就使得原先就不简单的研究问题变得更加复杂。因此，需要有一个将多个具有一定相关性的指标化为少数几个相互不相关的综合指标的统计方法来解决此类问题，这种方法被称为主成分分析，又称主分量分析。主成分概念最早由 Karl parson 于 1901 年在讨论非随机变量时引进，后经 Hotelling 于 1933 年将这一概念推广到随机向量，从而使得主成分分析在许多学科中都得到广泛应用。

为将多个(假设为 p 个)指标化为少数几个相互不相关的综合指标,在数学处理上,就是将原来的多个指标做线性组合, $F_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{pi}X_p$ 。在不加任何限制的条件,这种线性组合有很多,因此,需要对这种线性组合提出以下要求:

(2) 前面主成分的信息不再在后面的主成分中出现, 即各主成分间应相互独立;

12.1.2 主成分分析的数学模型及求法

$$X = (X_1, X_2, \dots, X_p)$$
$$X_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{bmatrix} \quad i = 1, 2, \dots, p$$
$$\left\{ \begin{array}{l} F_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p \\ F_2 = a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p \\ \quad \quad \quad \dots\dots\dots \\ F_p = a_{1p}X_1 + a_{2p}X_2 + \cdots + a_{pp}X_p \end{array} \right.$$
$$F_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{pi}X_p = \mathbf{a}'_i\mathbf{X} \quad i=1,2,\cdots,p$$

因此, 要求出主成分 F_1, F_2, \dots, F_p , 实际上就是要求上述方程组的系数矩阵 \mathbf{a}_i 的解。

当 \mathbf{a}_i 给定后, 对每个样品就可求出 F_i 值, 从而对 n 个样品计算 F_i 值就可得到 n 个 F_i 值, 由这 n 个 F_i 值可求出一个方差, 记为 $\text{Var}(F_i)$, 因此, 要求出系数矩阵 \mathbf{a}_i 的解, 只需要对主成分再做出一些条件限制即可, 首先各主成分间应互不相关, 另外, F_1 是 X_1, X_2, \dots, X_p 的一切线性组合中方差最大的, F_2 是与 F_1 不相关的 X_1, X_2, \dots, X_p 一切线性组合中方差最大的, \dots, F_p 是与 F_1, F_2, \dots, F_{p-1} 都不相关的 X_1, X_2, \dots, X_p 的一切线性组合中方差最大的。

在满足上述条件的基础上, 可以证明: 系数 a_{ij} 不是别的, 它恰好就是 \mathbf{X} 的协差阵 Σ 的特征值所对应的特征向量。而主成分 F_i 对应的特征值 λ_i 就是该主成分 F_i 的方差。

第 1 个主成分综合变量 (指标) X_1, X_2, \dots, X_p 信息的能力用其贡献率来表示, 它被定义为: $\lambda_1 / \sum_{i=1}^p \lambda_i = \lambda_1 / p$, 它就是第 1 主成分的方差在总方差中所占的比例, 该值越大, 表明其解释 X_1, X_2, \dots, X_p 信息的能力越强。

在解决实际问题时, 一般不需要取 p 个主成分, 而只需取累计贡献率达到 70%~80% 以上的前 k 个主成分即可。前 k 个主成分的累计贡献率通过下述公式计算

$$\text{前 } k \text{ 个主成分的累计贡献率} = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \lambda_i}, \quad i = 1, 2, \dots, p$$

当协差阵 Σ 未知时, 可用样本协差阵 \mathbf{S} 来估计,

$$\mathbf{S} = (\mathbf{s}_{ij})$$

其中

$$s_{ij} = \frac{1}{n} = \sum_{\alpha=1}^n (x_{\alpha i} - \bar{x}_i)(x_{\alpha j} - \bar{x}_j)$$

而相关系数阵

$$\mathbf{R} = (\mathbf{r}_{ij})$$

其中

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

显然, 当原始变量 X_1, X_2, \dots, X_p 标准化后, 则

$$\mathbf{S} = \mathbf{R} = \frac{1}{n} \mathbf{X}' \mathbf{X}$$

实际应用时, 由于指标的单位量纲不一, 所以在计算前应先对原始数据做标准化处

理, 以便消除量纲的影响。

12.1.3 主成分的性质

p 个主成分 (或称综合指标) 有以下几个简单性质:

- (1) 各个主成分互不相关;
- (2) 第 i 个主成分的方差等于其对应的特征值 λ_i , 即 $Var(F_i) = \lambda_i$;
- (3) 所有主成分的方差之和等于指标数, 即 $\lambda_1 + \lambda_2 + \cdots + \lambda_p = p$;
- (4) 第 i 个主成分 F_i 与变量 (指标) x_j 的相关系数为 $\sqrt{\lambda_i} a_{ji}$;

$$(5) \lambda_i = \sum_{j=1}^p a_{ij}^2 \lambda_j \quad i = 1, 2, \cdots, p。$$

12.1.4 主成分的应用及其注意点

(1) 从协差阵和相关系数阵出发, 计算出的主成分一般是不同的, 考虑到单位量纲的不同, 为便于对结果的解释, 建议先对原始数据进行标准化处理。

(2) 主成分是原始变量的线性组合, 它不能简单地解释为单变量的属性作用, 因而不能直接说明单个原始变量属性对主成分的作用, 而应同时看一些起主要作用的原始变量的综合作用, 以此给主成分一个合理的解释。

(3) 主成分分析并不要求分布为多元正态。

(4) 一般取前三个主成分, 当累计贡献率达到 70%~80% 以上即可。

(5) 主成分可用来建立评估系统及用它可做主成分回归。

(6) 要检验变量间的相关性。当各个变量间都不太相关时, 是无法进行降维的, 此时也只有变量自己代表自己了。因此, 可把变量之间是否有相关性, 当作能否进行主成分分析的前提条件来看。

对变量间的相关性检验, 在变量的相关系数矩阵作为研究的出发点的前提下, 有好几种检验方法, 如在第 6 章中介绍过的 Bartlett 球型检验, 此时, 当原始数据做标准化处理后, 数据资料服从均值为 0, 方差为 1 的标准正态分布, 所以, 此时其原假设为相关系数矩阵是一个单位矩阵, 即相关系数矩阵对角线上的所有元素都为 1, 而所有非对角线上的元素都为 0。由此可以用原先用来做随机误差独立性检验的 Bartlett 球型检验来检验原始变量间的相关性检验了。另外, 反映像相关矩阵检验 (Anti-image correlation matrix) 也是一种变量间相关性检验方法。它以变量的偏相关系数矩阵为出发点, 对每个偏相关系数矩阵的元素取倒数, 得反映像相关矩阵。如果变量间存在较多的重叠影响, 则变量间的偏相关系数较小, 因而在反映像相关矩阵中出现对应的那些元素的绝对值比较大, 说明这些变量不适合做主成分分析。在主成分分析中, 对变量间相关性检验用得较多的

方法是 KMO (Kaiser-Meyer-Olkin) 检验法, KMO 统计量的计算公式为

$$KMO = \frac{\sum_{i \neq j} \sum_{j \neq i} r_{ij}^2}{\sum_{i \neq j} \sum_{j \neq i} r_{ij}^2 + \sum_{i \neq j} \sum_{j \neq i} p_{ij}^2}$$

其中, r_{ij}^2 是变量 i 和变量 j 之间的简单相关系数, p_{ij}^2 是变量 i 和变量 j 之间的偏相关系数。

KMO 值越接近于 1, 说明所有变量之间的简单相关系数平方和远大于偏相关系数的平方和, 则可以做主成分分析。

Kaiser 给出了用 KMO 值来判断是否适合做主成分 (因子) 分析的标准:

0.9 < KMO: 非常适合;

0.8 < KMO < 0.9: 适合;

0.7 < KMO < 0.8: 一般;

0.6 < KMO < 0.7: 不太适合;

KMO < 0.6: 不适合。

因此, 在做主成分 (因子) 分析前, 可先用 KMO 检验法, 来判断一下所得到的诸多变量是否适合做主成分 (因子) 分析, 不失为一种明智的做法。

12.1.5 主成分实例分析

例 12.1 以例 10.7 为例, 试对随机抽取的某高中高一男生 38 人测试的反映其机能、体能和运动能力等 12 项指标 (原始数据见表 10-18) 进行主成分分析。数据资料已存放在 data12-01.sav 中。

在 SPSS 中的操作步骤如下:

(1) 在数据编辑窗口中, 打开 data12-01.sav。

(2) 在 Analyze 下拉式菜单 Data Reduction 中选择 Factor, 打开 Factor Analysis 对话框, 见图 12-1, 将所要分析的变量反复横向跳、纵跳、背力、握力、台阶试验、立定体前屈、俯卧上体后仰、450 米跑、跳远、投球、引体向上、耐力跑选中后, 用按钮移入右框。

(3) 单击 Descriptives 按钮, 进入 Factor Analysis: Descriptives 对话框, 见图 12-2。在 Correlation Matrix 选项中, 选择 KMO and Bartlett's test of sphericity 选项, 要求用 Bartlett 球形检验对变量的相关矩阵进行相关分析并计算偏相关矩阵的 KMO 统计量。

关闭本对话框中的其他选项, 单击 Continue 按钮返回 Factor Analysis 对话框 (见图 12-1)。

(4) 单击 Extraction 按钮, 打开 Factor Analysis: Extraction 对话框, 见图 12-3。在 Extraction 对话框中, 在 Method 下拉列表中选择 Principal Component, 要求进行主成分分析。

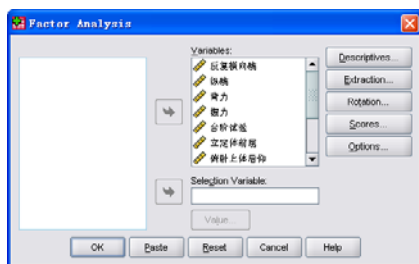


图 12-1 Factor Analysis 对话框

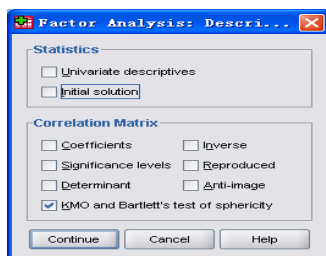


图 12-2 Descriptives 对话框

在 Analyze 选择项中选择 Correction matrix, 要求根据相关矩阵, 进行主成分分析。在 Display 选择项中选择 Unrotated factor solution 和 Scree plot, 要求输出未旋转因子解和碎石图。在 Extract 选择项中选择保持系统默认选项, 即选取特征值大于等于 1 的主成分。

单击 Continue 按钮, 返回 Factor Analysis 对话框 (见图 12-1)。

(5) 单击 Rotation 按钮, 打开 Factor Analysis: Rotation 对话框, 见图 12-4。

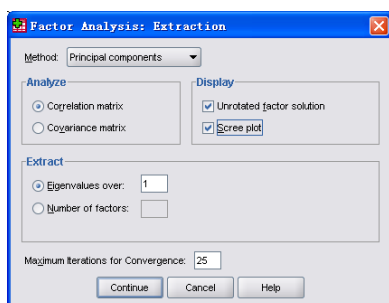


图 12-3 Factor Analysis: Extraction 对话框

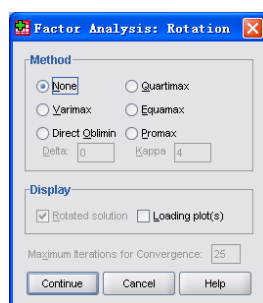


图 12-4 Factor Analysis: Rotation 对话框

在 Method 选择项中, 选择系统默认选项 None。

单击 Continue 按钮, 返回 Factor Analysis 对话框 (见图 12-1)。

其他采用系统默认值。

(6) 单击 OK 按钮执行, 在输出窗口中得到计算结果, 见表 12-1、表 12-2、表 12-3, 及图 12-5。表 12-4 不是在输出窗口中直接得到而是中间通过人工计算得到的表。

表 12-1 KMO 统计量和球型检验

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		
		.615
Bartlett's Test of Sphericity	Approx. Chi-Square	171.023
	df	66
	Sig.	.000

表 12-2 主成分特征值占总方差的比例

Component	Total Variance Explained		
	Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	4.052	33.770	33.770
2	1.753	14.604	48.374
3	1.453	12.108	60.482
4	1.224	10.201	70.683

Extraction Method: Principal Component Analysis.

(7) 结果与讨论

从表 12-1 中可以看到，球型检验的结果表明，在相关系数矩阵是一个单位矩阵的原假设下，观测的显著性水平为 0.000，故拒绝变量间全部相互独立的原假设，说明这些变量间至少有两个是相关的，也即它们之间有简单线性相关关系，可做主成分分析。但 KMO 的值为 0.615，较小，表明变量之间虽然存在相关，但相互独立的变量也较多，此时要使前几个主成分的累计贡献率达到 70% 以上，会有较多的主成分出现，表 12-2 是输出窗口中的最后一张表，该表列出了四个主成分对应的特征值、占总方差的百分比及累计百分比。第一主成分占总方差的百分比为 21.712%，说明用它可解释原来变量中 21.712% 的信息。前四个主成分的累计贡献率为 70.683%，说明用四个主成分来表示原来 12 个变量，可以反映 70.683% 的信息。故依 Kaiser 判断标准，应归结为本题不太适合做主成分分析。

但为了能理解 Kaiser 判断标准的合理性，及看到主成分分析的全过程，我们对本例结果继续往下做分析。

从图 12-5 可见，特征值大于 1 的主成分有 4 个，从第 4 个主成分开始图形出现拐点。因此，本例分 4 个主成分比较合适，这同表 12-3 中得到的结果相一致。

表 12-3 给出了 4 个主成分未做方差旋转时的因子载荷矩阵。

这里每一列代表一个主成分作为原来变量线性组合的系数。

同一个变量在 4 个主成分上，有不同的载荷（即它同各主成分的相关系数），相关系数越大，主成分对它的代表性越高。

在许多书籍中，都用表 12-4 中的因子载荷来作为各主成分表达式中的系数，而给出

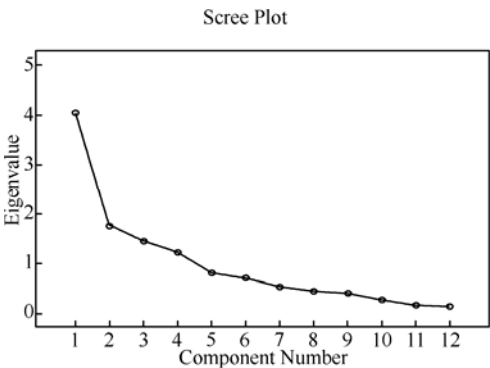


图 12-5 12 个成分的特征值的碎石图（点图）

表 12-3 未旋转因子载荷矩阵

	Component			
	1	2	3	4
反复横向往跳	.526	-.072	-.682	-.041
纵跳	.722	-.006	.183	.376
背力	.683	-.314	.250	-.268
握力	.309	.264	.391	-.662
台阶试验	.069	.711	-.553	.101
立定体前屈	.474	.073	.270	.581
俯卧上体后仰	.197	.708	.405	.266
50米跑	-.578	-.547	.228	.281
跳远	.793	.000	.238	-.125
投球	.722	-.286	-.130	.219
引体向上	.726	.143	-.038	-.093
耐力跑	-.642	.406	.264	.041

Extraction Method: Principal Component Analysis.
a. 4 components extracted.

表 12-4 主成分对应的特征向量表

	第一特征向量	第二特征向量	第三特征向量	第四特征向量
反复横向往跳	0.261261603	-0.054555328	-0.565791146	-0.036775456
纵跳	0.358721625	-0.004763482	0.151691985	0.339897242
背力	0.339438658	-0.236940064	0.207261322	-0.241898479
握力	0.153514132	0.199346721	0.324065212	-0.59848266
台阶试验	0.034043432	0.537354106	-0.458962536	0.091007494
立定体前屈	0.235581336	0.055441589	0.223650098	0.52482233
俯卧上体后仰	0.09761547	0.535155903	0.336036922	0.240844349
50米跑	-0.287247287	-0.412835493	0.189362097	0.253904305
跳远	0.393747537	-0.000402947	0.197311776	-0.112930095
投球	0.358588624	-0.215809313	-0.107764351	0.197616179
引体向上	0.360825195	0.108229204	-0.031579583	-0.084139893
耐力跑	-0.318999888	0.30689696	0.218877246	0.036632531

各主成分的表达式的。有的还用旋转后的因子载荷来给出主成分表达式。由于这样的情况实在太多，所以这里不一一列举，也不作过多的评述。事实上，根据主成分的定义可知，主成分表达式中的系数应是各特征值对应的特征向量。

特征向量与因子载荷之间有如下的关系

$$u_{ij} = \frac{r_{F_i x_j}}{\sqrt{\lambda_i}} \quad j = 1, 2, \dots, p$$

式中， $r_{F_i x_j}$ 表示因子载荷，即第 i 个变量与第 j 个主成分之间的相关系数， λ_i 为第 i 个主成分对应的特征值。

所以要求出特征向量的值，只需知道各主成分对应的特征值即可。各主成分对应的特征值可在表 12-3 找到。如第一主成分对应的特征值为 $\lambda_1 = 4.052$ ，第二主成分对应的特征值为 $\lambda_2 = 1.753$ 等。由此可得到各特征值对应的特征向量见表 12-5。

根据表 12-5 中的这些系数，我们可以写出各个主成分具体的表达式，4 个主成分分别为

$$\begin{aligned} F_1 &= 0.2613x_1 + 0.3587x_2 + 0.3394x_3 + 0.1535x_4 + 0.0340x_5 + 0.2356x_6 \\ &\quad + 0.0976x_7 - 0.2872x_8 + 0.3937x_9 + 0.3586x_{10} + 0.3608x_{11} - 0.3190x_{12} \\ F_2 &= -0.0546x_1 - 0.0048x_2 - 0.2369x_3 + 0.1993x_4 + 0.5373x_5 + 0.0554x_6 \\ &\quad + 0.5352x_7 - 0.4128x_8 - 0.0004x_9 - 0.2158x_{10} + 0.1082x_{11} + 0.3069x_{12} \\ F_3 &= -0.5658x_1 + 0.1517x_2 + 0.2073x_3 + 0.3241x_4 - 0.4590x_5 + 0.2237x_6 \\ &\quad + 0.3360x_7 + 0.1894x_8 + 0.1973x_9 - 0.1078x_{10} - 0.0316x_{11} + 0.2189x_{12} \\ F_4 &= -0.0368x_1 + 0.3399x_2 - 0.2419x_3 - 0.5985x_4 + 0.0910x_5 + 0.5248x_6 \\ &\quad + 0.2408x_7 + 0.2539x_8 - 0.1129x_9 + 0.1976x_{10} - 0.0841x_{11} + 0.0366x_{12} \end{aligned}$$

式中各代码的含义为： x_1 —反复横向跳（次）、 x_2 —纵跳（cm）、 x_3 —背力（Kg）、 x_4 —握力（Kg）、 x_5 —台阶试验（指数）、 x_6 —立定体前屈（cm）、 x_7 —俯卧上体后仰（cm）（俯卧背伸测验）、 x_8 —50 米跑（秒）、 x_9 —跳远（cm）、 x_{10} —投球（m）、 x_{11} —引体向上（次）、 x_{12} —耐力跑（秒）。

主成分的命名由各主成分中起主要作用的那些变量所代表的属性来决定。变量对哪个主成分起主要作用，由同一变量在各主成分中特征向量绝对值的大小来决定。

在第一主成分中，第 2 项指标、第 3 项指标、第 9、10、11 项和第 12 项指标的绝对值较大，说明这六个指标起主要作用，因此，可把第一主成分看成是爆发力综合指标。

在第二主成分中，第 5 项指标、第 7 项指标和第 8 项指标的绝对值较大，说明这三个指标起主要作用，鉴于第 5 项指标的影响最大又与第 8 项指标成负关系（它是反映无氧工作能力的），因此，可把第二主成分看成是心血管耐力综合指标。

在第三主成分中，第 1 项指标的绝对值较大，说明这个指标起主要作用，因此，可把第三主成分看成是灵敏性综合指标。

在第四主成分中,第 4 项指标、第 6 项指标的绝对值较大,说明这二个指标起主要作用,考虑到第 6 项指标的系数为正,而第 4 项指标的系数为负,因此,可把第四主成分看成是后侧肌群的伸展能力综合指标。

以上分类命名是建立在体育测量学的专业基础知识上给出的,正如在表 12-2 中看到的一样,KMO 的值为 0.615,较小,表明变量之间虽然存在相关,但相互独立的变量也较多,因此,这也给取较多主成分时的命名工作带来困难。因此,以上对各主成分的命名不一定是最佳的,仅作参考。

输出窗口中,倒数第二张表是反映共同度的,在因子分析中才用到,由于它与主成分分析无关,所以关于共同度的具体统计意义,这里先不作进一步解释,也不列出该表。

由本例可知,当 KMO 的值低于 0.70 时,由于变量之间相关性不高,要解释原来变量中 70% 以上的信息时,需要动用较多的主成分,这无疑没有达到通过主成分分析来降维的目的。一般取前三个主成分已经足够多了,所以要求变量之间应尽可能相关。

12.2 因子分析

我们知道,随着年龄的增加,儿童的身高、体重也会随之变化,并体现出相关性,变量之间存在相关性的原因是由于它们都受到了一些共同因子(共性因子)的支配。儿童的身高、体重之所以能同时变化,就是由于它们同时受到生长因子的支配。因此,因子分析的任务就是要从大量的数据中寻找影响、支配变量的更本质的因子——共性因子。这种共性因子有时不止一个而是多个。

主成分分析是因子分析的一个特例,因子分析是主成分分析的推广和发展。它也是一种多元统计分析中处理降维的一种统计方法。最早用于心理学和教育学方面的研究,在 Charles Spearman 发表的论文《对智力测验得分进行统计分析》(1904)中,首次出现因子分析的提法。现广泛应用于自然科学和社会科学的各个领域。

12.2.1 因子分析的数学模型及模型系数的统计意义

1. 因子分析的数学模型

因子分析有很多方法,常用的有 R 型因子分析和 Q 型因子分析两种方法。R 型因子分析和 Q 型因子分析从计算过程来看,两者是一样的,用的都是同一批观测数据,所不同的是,R 型因子分析使用的是变量的相关矩阵,而 Q 型因子分析使用的是样品的相似系数矩阵。因此,这里主要讨论 R 型因子分析。

设对每个样品观测 p 个相互间有相关性的指标 X_1, X_2, \dots, X_p , 共观测 n 个样品,则 p 个指标(变量)组成的向量 $X = (X_1, X_2, \dots, X_p)$ 。

假设,这些变量都已事先进行了标准化处理,则每个变量的样本均值都为 0,方差也

都为 1。

由于变量间相关性的存在，因此，在理论上，总可以将每个变量分解成两个部分

$$X_i = X_i^* + \varepsilon_i \quad (12.1)$$

其中 X_i^* 是 X_i 中与其他变量相关的部分， ε_i 是 X_i 中与其他变量不相关的部分。

如果 X_i 与其他任何变量都没有相关性，则 $X_i^* = 0$ ， $X_i = \varepsilon_i$ ，反之，如果 X_i 可以用其他变量的线性组合来表示，则 $\varepsilon_i = 0$ 。

假设支配 p 个变量的公共因子不止一个而有 m 个 ($m \leq p$)，记为 F_1, F_2, \dots, F_m ， F_1, F_2, \dots, F_m 彼此不相关且方差为 1， $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$ 互不相关，且方差不同，分别为 $\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2$ 。则式 (12.1) 可以表示成下面的形式

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \dots + a_{1m}F_m + \varepsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \dots + a_{2m}F_m + \varepsilon_2 \\ \vdots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \dots + a_{pm}F_m + \varepsilon_p \end{cases} \quad (12.2)$$

它是 R 型因子分析的数学模型。其中， a_{ij} 表示 X_i 在公共因子 F_j 上的载荷系数（也称权重系数）， $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)$ 是与 (F_1, F_2, \dots, F_m) 不相关的因子，称特殊因子（也称唯一性因子）。

R 型因子分析的数学模型用矩阵可表示为

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

因而它可简记为

$$X = AF + \varepsilon$$

由上可知， $X_i^* = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{im}F_m \quad i = 1, 2, \dots, p \quad (12.3)$

在上述模型中，共性因子 (F_1, F_2, \dots, F_m) 有两种情况，一是共性因子 (F_1, F_2, \dots, F_m) 彼此间不相关，称它为正交模型，二是共性因子 (F_1, F_2, \dots, F_m) 之间可以有相关性，称它为斜交模型。由于在斜交模型中，对参数的解释要比正交模型下复杂得多，因此，这里只讨论公共因子间不相关的情况。

2. 因子载荷 a_{ij} 的统计意义

在每个变量、每个公共因子和唯一性（特殊）因子都已标准化下，理论上可以证明，因子载荷 a_{ij} 就是第 i 个变量与第 j 个公共因子的相关系数，它反映了变量 X_i 在 F_j 上的相

对重要性。

3. 变量共同度的统计定义及统计意义

变量共同度,也称为公共方差。变量 X_i 的共同度定义为因子载荷矩阵 \mathbf{A} 中第 i 行元素的平方和,即

$$h_i^2 = a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 = \sum_{j=1}^m a_{ij}^2 \quad i=1,2,\cdots,p$$

由于 X_i 已经标准化,所以,可以证明

$$1 = h_i^2 + \sigma_i^2$$

此式表明,变量 X_i 的方差由两部分组成,第一部分为共同度 h_i^2 ,它反映公共因子对变量 X_i 的总方差的解释的比例(或所作的贡献), h_i^2 越接近于 1,说明公共因子解释了变量 X_i 越多的信息。当 $h_i^2 \approx 0$ 时,说明公共因子对 X_i 影响很少,主要由特殊因子 ε_i 来描述。第二部分 σ_i^2 是特定变量所产生的方差,仅与 X_i 本身的变化有关,称为特殊因子方差,它是使 X_i 的方差为 1 的补充值。

4. 公共因子 F_j 的方差贡献的统计意义

公共因子 F_j 的方差贡献定义为因子载荷矩阵中各列元素的平方和,其计算公式为

$$S_j = \sum_{i=1}^p a_{ij}^2 \quad j=1,2,\cdots,p$$

它是衡量公共因子相对重要性的指标。其值越高,说明该因子的重要程度越高。

12.2.2 因子载荷矩阵的估计

建立因子模型的关键是要根据样本数据矩阵估计因子载荷矩阵 \mathbf{A} 。对它的估计有简易估计法、精细估计法、主因子解法和最大似然估计法等很多方法,这里介绍使用较为普遍的主成分法。

设随机向量 $X = (X_1, X_2, \cdots, X_p)$ 的协差矩阵为 Σ , $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ 为其特征根, u_1, u_2, \cdots, u_p 为其对应的标准正交化特征向量。

在特殊因子的方差为 0 时,因子分析的模型为 $X = AF$, 其中 $D(F) = I_m$ (因子的方差矩阵为单位矩阵)。因此, $D(X) = D(AF) = AD(F)A' = AA' = \Sigma$, 而由线性代数的知识可知

$$\Sigma = U \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix} U' = \sum_{i=1}^p \lambda_i u_i u_i' = \left(\sqrt{\lambda_1} u_1, \cdots, \sqrt{\lambda_p} u_p \right) \begin{bmatrix} \sqrt{\lambda_1} u_1' \\ \vdots \\ \sqrt{\lambda_p} u_p' \end{bmatrix}$$

由上可知, 因子载荷矩阵 A 的第 j 列为 $\sqrt{\lambda_j}u_j$, 这也就是说, 除常数 $\sqrt{\lambda_j}$ 外, 第 j 列的因子载荷正是第 j 列的主成分的系数 u_j , 所以称这样的因子载荷的估计方法为主成分法。

由于在实际应用中, 总是希望公共因子数 m 小于变量的个数 p , 因此, 当最后 $p-m$ 个特征根较小时, 最后 $p-m$ 项 $\lambda_{m+1}u_{m+1}u_{m+1}' + \cdots + \lambda_p u_p u_p'$ 对 Σ 的贡献总是被省略。

在考虑特殊因子时, 此时的协差阵为

$$\Sigma \approx AA' + \Sigma_\varepsilon = \left(\sqrt{\lambda_1}u_1, \dots, \sqrt{\lambda_m}u_m \right) \begin{bmatrix} \sqrt{\lambda_1}u_1' \\ \vdots \\ \sqrt{\lambda_m}u_m' \end{bmatrix} + \begin{pmatrix} v_1^2 & & 0 \\ & \ddots & \\ 0 & & v_p^2 \end{pmatrix}$$

由于 Σ 通常是未知的, 此时, 一般用样本协差阵 S 来代替, 为消除单位量纲的影响, 很多时候都要对变量进行标准化处理, 则此时的样本协差阵 S 与样本相关阵 R 相同, 故依然可以用上式表示。

设 $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ 为样本相关阵的特征根, $\hat{u}_1, \hat{u}_2, \dots, \hat{u}_p$ 为其对应的标准正交化特征向量。在公共因子数 m 小于变量的个数 p 时, 因子载荷阵的估计 $\hat{A} = (\hat{a}_{ij})$, 也就是

$$\hat{A} = (\sqrt{\hat{\lambda}_1}\hat{u}_1, \dots, \sqrt{\hat{\lambda}_m}\hat{u}_m)$$

12.2.3 因子旋转

为便于对实际问题进行科学的分析, 必须明确每个公共因子的实际意义。满足模型式 (12.2) 中的 A 阵可以有无限组, 记 C 为任意一个 m 阶正交阵, 并令 $B=AC$, 显然有

$$BB' = ACC'A' = AA' \quad (12.4)$$

它说明因子载荷阵不是唯一的。正因为此, 可以通过变换因子载荷矩阵使因子载荷结构更加简化, 也就是使每个变量仅在一个公共因子上有较大的载荷, 而在其他公共因子只有较小或中等的载荷, 这对公共因子的解释无疑是有益的。这种对因子载荷阵进行变换的方法称为因子轴的旋转。因子轴旋转的方法很多, 其中最有名的是 Kaiser 于 1951 年提出的方差最大正交旋转。这里只对此法进行介绍。

在式 (12.4) 中寻找 B 载荷阵的方法, 相当于把度量 A 元素的坐标轴作了一次旋转, 也就是将原来的坐标 A 变为新坐标轴下的 B 。对 m 个因子同时旋转相当于对每两个因子作旋转, 故不妨先讨论公共因子数 $m=2$ 的情形。

$$\text{设因子载荷阵 } A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ \vdots & \vdots \\ a_{p1} & a_{p2} \end{bmatrix}, \text{ 正交阵 } C = \begin{pmatrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{pmatrix}$$

$$\text{则 } B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ \vdots & \vdots \\ b_{p1} & b_{p2} \end{bmatrix} = AC = \begin{bmatrix} a_{11} \cos \varphi + a_{12} \sin \varphi & -a_{11} \sin \varphi + a_{12} \cos \varphi \\ a_{21} \cos \varphi + a_{22} \sin \varphi & -a_{21} \sin \varphi + a_{22} \cos \varphi \\ \vdots & \vdots \\ a_{p1} \cos \varphi + a_{p2} \sin \varphi & -a_{p1} \sin \varphi + a_{p2} \cos \varphi \end{bmatrix}$$

旋转的目的是使因子载荷矩阵 A 的结构简化, 即使公共因子的贡献越分散越好, 也就是正交旋转的角度 φ 须满足使旋转后所得到的因子载荷阵的总方差 V 达到最大值。

因此, 要使 $V = V_1 + V_2 = \frac{1}{p} \sum_{j=1}^2 \sum_{i=1}^p \left(\frac{b_{ij}^2}{h_i^2} \right) - \frac{1}{p^2} \sum_{j=1}^2 \left(\sum_{i=1}^p \frac{b_{ij}^2}{h_i^2} \right)^2$ 达到最大 (式中, h_i^2 为共同度,

$h_i^2 = \sum_{j=1}^2 a_{ij}^2, i=1, 2, \dots, p$, b_{ij} 为待定系数), 根据极值原理可知, 它等同于求 V 对 φ 的导数

并令其等于 0, 得: $\frac{\partial V}{\partial \varphi} = \omega \cos 4\varphi - \delta \sin 4\varphi = 0$

因此 $\text{tg } 4\varphi = \frac{\omega}{\delta}$ (12.5)

其中, $\delta = Y - (W^2 - X^2)/p$, $\omega = Z - 2WX/p$, 而 $W = \sum_{i=1}^p Q_i$, $X = \sum_{i=1}^p E_i$, $Y = \sum_{i=1}^p (Q_i^2 - E_i^2)$,

$Z = 2 \sum_{i=1}^p Q_i E_i$, $Q_i = (a_{i1}/h_i)^2 - (a_{i2}/h_i)^2$, $E_i = 2a_{i1} a_{i2}/h_i^2$ 。

根据 $\text{tg } 4\varphi = \frac{\omega}{\delta}$ 的分子和分母取值的正负号可以确定 φ 的取值范围, 见表 12-5。

表 12-5 φ 的取值范围

分子取值符号	分母取值符号	φ 的取值范围
+	+	$0 \sim \frac{\pi}{8}$
+	-	$\frac{\pi}{8} \sim \frac{\pi}{4}$
-	-	$-\frac{\pi}{4} \sim -\frac{\pi}{8}$
-	+	$-\frac{\pi}{8} \sim 0$

当公共因子数为 m 个时, 则需逐次对每两个公共因子进行上述旋转, 它共需旋转 $\frac{m(m-1)}{2}$ 次, 算作一个循环完毕, 此时的因子载荷矩阵为

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix}$$

需要旋转的因子为 F_k , F_j , 所以, 此时的正交阵

$$C_{kj} = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & \cos \varphi & & -\sin \varphi \\ & & & & 1 & \\ & & & & & \ddots \\ & & & & & & 1 \\ & & \sin \varphi & & & \cos \varphi & \\ & & & & & & 1 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{bmatrix}$$

矩阵中, 第 k, j 列和第 k, j 行交叉处, 为如上所示的余弦和正弦函数, 除此以外, 其他非对角线上的元素为 0, 对角线上的元素为 1。

$$\text{则 } B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{p1} & b_{p2} & \cdots & b_{pm} \end{bmatrix} = AC_{kj}$$

其元素为

$$\begin{aligned} b_{ik} &= a_{ik} \cos \varphi + a_{ij} \sin \varphi \\ b_{ij} &= -a_{ik} \sin \varphi + a_{ij} \cos \varphi \quad i = 1, 2, \dots, p \\ b_{il} &= a_{il} \quad (l \neq k, j) \end{aligned}$$

在 m 个因子时, B 的总的相对方差为

$$V = \sum_{j=1}^m V_j = \frac{1}{p} \sum_{j=1}^m \sum_{i=1}^p \left(\frac{b_{ij}^2}{h_i^2} \right) - \frac{1}{p^2} \sum_{j=1}^m \left(\sum_{i=1}^p \frac{b_{ij}^2}{h_i^2} \right)^2 \quad (12.6)$$

其中 $h_i^2 = \sum_{j=1}^m a_{ij}^2$, b_{ij} 为待定系数。旋转角度 φ 仍按上面 (12.5) 式求得。

记第一轮旋转循环完毕的因子载荷矩阵为 $B_{(1)}$, 用式 (12.6) 来算得旋转后的因子载

荷的方差为 $V_{(1)}$ ，在第一轮旋转循环完毕的基础上，从 $B_{(1)}$ 出发进行第二轮旋转循环，旋转完毕得 $B_{(2)}$ 和 $V_{(2)}$ ，依次得到 $B_{(3)}$ 和 $V_{(3)}$ ， \cdots 。

如此不断重复上述做法，则可得 V 值的一个非降序列

$$V_{(1)} \leq V_{(2)} \leq V_{(3)} \leq \cdots$$

由于因子载荷的绝对值小于等于 1，所以这个序列有上界，记这个序列的极限为 $\tilde{V} = \max\{V_{(i)}\}$ ，当循环次数 k 充分大时，如果能满足 $|V_{(k)} - \tilde{V}| < \varepsilon$ ， ε 为事先给定的研究精度，则停止旋转循环。最后得到的 $B_{(k)}$ ，即为旋转后的因子载荷矩阵。

12.2.4 因子得分

由于公共因子能反映与原始变量间的相关关系，因此，用公共因子来代表原始变量时，有时更有利于描述研究对象的特征。而将公共因子表示为变量（或样品）的线性组合

$$F_j = \beta_{j1}X_1 + \beta_{j2}X_2 + \cdots + \beta_{jp}X_p \quad j=1,2,\cdots,m$$

称为因子得分函数。用它可对样品进行分类等。

估计因子得分的方法有很多，如回归法、加权最小二乘法、标准化残差平方和法等。这里只介绍汤姆森回归法，它由 Thomson 于 1939 年提出。

在变量已作标准化处理的前提下，Thomson 提出的用公共因子对 p 个变量作回归的方程为

$$\hat{F}_j = b_{j1}X_1 + b_{j2}X_2 + \cdots + b_{jp}X_p \quad j=1,2,\cdots,m$$

$$\text{用样本值可得} \quad A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix}$$

$$\text{根据因子载荷的意义可得} \quad \begin{cases} b_{j1}r_{11} + b_{j2}r_{12} + \cdots + b_{jp}r_{1p} = a_{1j} \\ \vdots \\ b_{j1}r_{p1} + b_{j2}r_{p2} + \cdots + b_{jp}r_{pp} = a_{pj} \end{cases} \quad j=1,2,\cdots,m$$

$$\Leftrightarrow Rb_j = a_j \quad j=1,2,\cdots,m$$

$$\text{其中} \quad b_j = (b_{j1}, b_{j2}, \cdots, b_{jp})', \quad a_j = (a_{1j}, a_{2j}, \cdots, a_{pj})'$$

$$\text{因此} \quad b_j = R^{-1}a_j \quad j=1,2,\cdots,m$$

记

$$B = \begin{bmatrix} b'_1 \\ b'_2 \\ \vdots \\ b'_m \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{p1} & b_{p2} & \cdots & b_{pm} \end{bmatrix}$$

则

$$B = \begin{bmatrix} (R^{-1}a_1)' \\ \vdots \\ (R^{-1}a_m)' \end{bmatrix} = \begin{bmatrix} a'_1 \\ \vdots \\ a'_m \end{bmatrix} R^{-1} = A'R^{-1}$$

所以估计因子得分的计算公式为： $\hat{F} = \begin{bmatrix} \hat{F}_1 \\ \vdots \\ \hat{F}_m \end{bmatrix} = BX = A'R^{-1}X$

其中 $X = (X_1, X_2, \dots, X_p)$ 。

12.2.5 实例分析

例 12.2 试对全国 30 个省、市、自治区经济发展基本情况的八项指标作因子分析，原始数据见表 12-6，数据已存放在 data12-02.sav 中。

表 12-6 全国 30 个省、市、自治区经济发展基本情况的八项指标统计

省份	GDP	居民消费水平	固定资产投资	职工平均工资	货物周转量	居民消费价格指数	商品零售价格指数	工业总产值
北京	1394.89	2505	519.01	8144	373.9	117.3	112.6	843.43
天津	920.11	2720	345.46	6501	342.8	115.2	110.6	582.51
河北	2849.52	1258	704.87	4839	2033.3	115.2	115.8	1234.85
山西	1092.48	1250	290.9	4721	717.3	116.9	115.6	697.25
内蒙	832.88	1387	250.23	4134	781.7	117.5	116.8	419.39
辽宁	2793.37	2397	387.99	4911	1371.1	116.1	114	1840.55
吉林	1129.2	1872	320.45	4430	497.4	115.2	114.2	726.47
黑龙江	2014.53	2343	435.73	4145	824.8	116.1	114.3	1240.37
上海	2462.57	5343	996.48	9279	207.4	118.7	113	1642.95
江苏	5155.25	1926	1434.95	5943	1025.5	115.8	114.3	2026.64
浙江	3524.79	2249	1006.39	6619	754.4	116.6	113.5	916.59
安徽	2003.58	1254	474	4609	908.3	114.87	112.7	824.14
福建	2160.52	2320	553.97	5857	609.3	115.2	114.4	433.67
江西	1205.11	1182	282.84	4211	411.7	116.9	115.9	571.84
山东	5002.34	1527	1229.55	5145	1196.6	117.6	114.2	2207.69

(续表)

省份	GDP	居民消费水平	固定资产投资	职工平均工资	货物周转量	居民消费价格指数	商品零售价格指数	工业总产值
河南	3002.74	1034	670.35	4344	1574.4	116.5	114.9	1367.92
湖北	2391.42	1527	571.68	4685	849	120	116.6	1220.72
湖南	2195.7	1408	422.61	4797	1011.8	119	115.5	843.83
广东	5381.72	2699	1639.83	8250	656.5	114	111.6	1396.35
广西	1606.15	1314	382.59	5105	556	118.4	116.4	554.97
海南	364.17	1814	198.35	5340	232.1	113.5	111.3	64.33
四川	3534	1261	822.54	4645	902.3	118.5	117	1431.81
贵州	630.07	942	150.84	4475	301.1	121.4	117.2	324.72
云南	1206.68	1261	334	5149	310.4	121.3	118.1	716.65
西藏	55.98	1110	17.87	7382	4.2	117.3	114.9	5.57
陕西	1000.03	1208	300.27	4396	500.9	119	117	600.98
甘肃	553.35	1007	114.81	5493	507	119.8	116.5	468.79
青海	165.31	1445	47.76	5753	61.6	118	116.3	105.8
宁夏	169.75	1355	61.98	5079	121.8	117.1	115.3	114.4
新疆	834.57	1469	376.95	5348	339	119.7	116.7	428.76

数据来源：1996 年《中国统计年鉴》。

在 SPSS 中的操作步骤如下：

(1) 在数据编辑窗口中，打开 data12-02.sav。

(2) 在 Analyze 下拉式菜单 Data Reduction 中选择 Factor，打开 Factor Analysis 对话框，见图 12-1，将所要分析的变量 GDP、居民消费水平、固定资产投资、职工平均工资、货物周转量、居民消费价格指数、商品零售价格指数、工业总产值选中后，用按钮移入右框。

(3) 单击 Descriptives 按钮，进入 Factor Analysis: Descriptives 对话框，见图 12-2。在 Correlation Matrix 选项中，选择 KMO and Bartlett's test of sphericity 选项，要求用 Bartlett 球形检验对变量的相关矩阵进行相关分析并计算偏相关矩阵的 KMO 统计量。

关闭本对话框中的其他选项，单击 Continue 按钮返回 Factor Analysis 对话框（见图 12-1）。

(4) 单击 Extraction 按钮，打开 Factor Analysis: Extraction 对话框，见图 12-3。在 Extraction 对话框中，在 Method 下拉列表中选择 Principal Component，要求进行主成分分析。

在 Analyze 选择项中选择 Correctation matrix，要求从相关矩阵出发，进行主成分分析。在 Display 选择项中选择 Unrotated factor solution 和 Scree plot，要求输出未旋转因子解和碎石图。在 Extract 选择项中选择保持系统默认选项，即选取特征值大于等于 1 的主

成分。

单击 Continue 按钮，返回 Factor Analysis 对话框（见图 12-1）。

（5）单击 Rotation 按钮，打开 Factor Analysis: Rotation 对话框，见图 12-4。

在 Method 选择项中，选择 Varimax 选项，要求做最大方差旋转，以便在对主成分命名困难时，能对主成分进行命名。在 Display 中选择 Rotated solution 选项，要求在输出窗口中显示旋转解。其他采用系统默认值。

单击 Continue 按钮，返回 Factor Analysis 对话框（见图 12-1）。

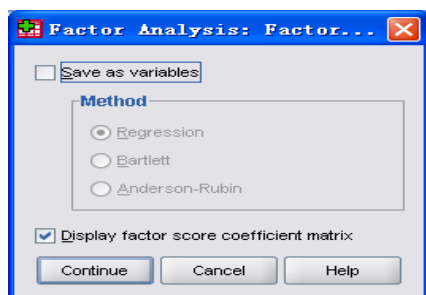


图 12-6 Factor Analysis: Factor...对话框
及图 12-7。

（8）结果与讨论。

从表 12-7 中可以看到，KMO 的值为 0.618，渐近卡方值为 231.889，自由度为 28，球型检验的结果表明，在相关系数矩阵是一个单位矩阵的原假设下，观测的显著性水平为 0.000，故拒绝原假设，说明这些变量各自不全独立，它们之间有简单线性相关关系，可做因子分析。但 KMO 的值较小，表明需要取较多的公共因子数才能使累计贡献率达到一定的要求。

（6）单击 Scores 按钮，打开 Factor Analysis: Factor...对话框，见图 12-6。

选择 Display factor score coefficient matrix 选项，要求在输出窗口中输出因子得分系数矩阵。

单击 Continue 按钮，返回 Factor Analysis 对话框（图 12-1）。

（7）单击 OK 按钮执行，在输出窗口中得到计算结果，见表 12-7 至表 12-14，

表 12-7 KMO 统计量和球型检验

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.618
Bartlett's Test of Sphericity	Approx. Chi-Square	231.889
	df	28
	Sig.	.000

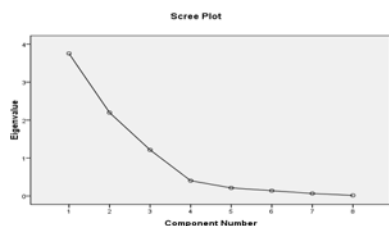


图 12-7 特征值碎石图

一定的要求。

从图 12-7 中可见，特征值大于 1 的共有三个，拐点出现在 4 处，它预示着本例中可能要用三个公共因子来进行分析时，将能概括变量内在的极大部分的信息。

表 12-8 给出了前三个公共因子未旋转和旋转后的特征值、占总方差的比例及累计贡献。由该表可知，前三个公共因子的累计贡献为 89.588%。这同图 12-7 中得到的结论相同。

表 12-8 贡献率

Total Variance Explained

Component	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.756	46.946	46.946	3.207	40.091	40.091
2	2.197	27.463	74.409	2.217	27.713	67.804
3	1.214	15.179	89.588	1.743	21.784	89.588

Extraction Method: Principal Component Analysis.

表 12-9 列出的是未旋转状态下的因子载荷矩阵, 由于居民消费水平在第一因子和第二因子上都有高载荷且比较相近, 及商品零售价格指数在第一因子和第二因子上也存在类似情况, 因此, 需要对此进行最大方差旋转, 以便更好地对因子进行命名。

表 12-9 未旋转因子载荷矩阵

Component Matrix^a

	Component		
	1	2	3
GDP	.885	.384	.120
居民消费水平	.607	-.598	.271
固定资产投资	.912	.161	.211
职工平均工资	.467	-.722	.369
货物周转量	.486	.738	-.277
居民消费价格指数	-.508	.254	.796
商品零售价格指数	-.619	.595	.437
工业总产值	.823	.427	.212

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

表 12-10 共同度

Communalities

	Extraction
GDP	.945
居民消费水平	.799
固定资产投资	.902
职工平均工资	.874
货物周转量	.857
居民消费价格指数	.957
商品零售价格指数	.928
工业总产值	.905

Extraction Method: Principal Component Analysis.

表 12-10 显示了三个因子对各指标的共同度, 除居民消费水平的共同度较低接近 80% 外, 其他指标上的共同度都达到 85% 以上, 表明用这三个公共因子可以反映原变量中绝大部分的信息。

表 12-11 将因子实行最大方差旋转后的载荷矩阵

Rotated Component Matrix^a

	Component		
	1	2	3
GDP	.955	.125	-.131
居民消费水平	.217	.841	-.214
固定资产投资	.871	.352	-.138
职工平均工资	.052	.927	-.115
货物周转量	.752	-.505	-.189
居民消费价格指数	-.135	-.010	.969
商品零售价格指数	-.102	-.494	.821
工业总产值	.944	.112	-.013

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

表 12-12 因子转换矩阵

Component Transformation Matrix

Co...	1	2	3
1	.817	.408	-.407
2	.548	-.768	.331
3	.178	.493	.851

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

表 12-11 给出了对因子实行最大方差旋转后得到的载荷矩阵, 从各因子上的高载荷可见, 在第一因子上, GDP、工业总产值、固定资产投资及货物周转量有高载荷, 因此, 可将第一因子命名为总量因子。在第二因子上, 职工平均工资、居民消费水平有高载荷, 因此, 可将第二因子命名为收支因子。在第三因子上, 居民消费价格指数、商品零售价

格指数有高载荷, 因此, 可将第三因子命名为价格因子。

我们用 x_1 、 x_2 、 x_3 、 x_4 、 x_5 、 x_6 、 x_7 、 x_8 分别表示 GDP、居民消费水平、固定资产投资、职工平均工资、货物周转量、居民消费价格指数、商品零售价格指数和工业总产值变量。这样因子 f_1 、 f_2 和 f_3 与这些变量之间的关系可用以下等式表示

$$x_1 = 0.955f_1 + 0.125f_2 - 0.131f_3$$

$$x_2 = 0.217f_1 + 0.841f_2 - 0.214f_3$$

$$x_3 = 0.871f_1 + 0.352f_2 - 0.138f_3$$

$$x_4 = 0.052f_1 + 0.927f_2 - 0.115f_3$$

$$x_5 = 0.752f_1 - 0.505f_2 - 0.189f_3$$

$$x_6 = -0.135f_1 - 0.010f_2 + 0.969f_3$$

$$x_7 = -0.102f_1 - 0.494f_2 + 0.821f_3$$

$$x_8 = 0.944f_1 + 0.112f_2 - 0.013f_3$$

表 12-12 列出了因子转换矩阵, 用来说明旋转前后因子间系数的对应关系, 据此来对因子进行相互转换。

表 12-13 因子得分系数矩阵

Component Score Coefficient Matrix			
	Component		
	1	2	3
GDP	.306	.010	.046
居民消费水平	.023	.385	.035
固定资产投资	.270	.128	.073
职工平均工资	-.024	.453	.099
货物周转量	.249	-.318	-.136
居民消费价格指数	.069	.179	.651
商品零售价格指数	.078	-.098	.463
工业总产值	.317	.026	.124

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

表 12-13 列出了因子得分的系数矩阵, 由此可以写出因子得分方程为

$$f_1 = 0.306x_1 + 0.023x_2 + 0.270x_3 - 0.024x_4 + 0.249x_5 + 0.069x_6 + 0.078x_7 + 0.317x_8$$

$$f_2 = 0.010x_1 + 0.385x_2 + 0.128x_3 + 0.453x_4 - 0.318x_5 + 0.179x_6 - 0.098x_7 + 0.026x_8$$

$$f_3 = 0.046x_1 + 0.035x_2 + 0.073x_3 + 0.099x_4 - 0.136x_5 + 0.651x_6 + 0.463x_7 + 0.124x_8$$

如果在 Factor Analysis: Factor...对话框 (图 12-6) 中选择 Save as variables 则由上面三式计算的三个因子得分将作为新变量存放在工作的数据文件中。

表 12-14 给出了因子得分的协方差矩阵, 由此可见, 各因子之间是正交的, 即相互之间彼此独立。

表 12-14 因子得分协方差矩阵

Component Score Covariance Matrix			
Component	1	2	3
1	1.000	.000	.000
2	.000	1.000	.000
3	.000	.000	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

12.3 对应分析

如果把数据资料看成由行、列变量组成的矩阵,那么,在多种处理降维的统计方法中,第 10 章中见到的典型相关分析,就是研究列中两组变量之间关系的一个统计分析方法,而在此之前提到的 R 型因子分析(或主成分分析)就是对变量(列中的变量)进行降维处理的一种统计方法,而 Q 型因子分析(或主成分分析)就是对样品(观测值或行中的变量)进行降维处理的一种统计方法。

在前面的学习中,我们已经知道 R 型因子分析和 Q 型因子分析在计算的数学原理上没有什么本质的区别,但这两个过程是分开进行的。显然,这样做的结果,一方面会漏掉一些指标和样品间的信息,另一方面因因子分析要求观测量(样品)数目必须是变量(指标)数的 5 倍,因此,在做 Q 型因子分析时,还要做比 R 型因子分析计算量更大的计算工作。这从研究设计的要求来说,并不是最佳的。

在实际研究中,研究人员很多情况下所关心的除行或列本身变量之间的关系外,更想了解行变量和列变量的相互关系。因此,改良算法,避免计算过程的重复,将 R 型和 Q 型因子分析在一次计算中同时完成,从而达到总计算量最小而又同时考虑指标和样品的关系,是有很重要的实际意义的。

这种通过改良算法,在一次计算中,将 R 型和 Q 型因子分析同时完成的统计方法称为对应分析,也称为相应分析。它首先由 M.Richardson 和 G.F.Kuder 于 1933 年提出,之后法国统计学家 J.P.Beozecri 和日本统计学家林知己夫对它的理论和方法进行了深入的研究。

对应分析不但可以处理连续变量的数据矩阵,也可应用于列联表数据的统计分析。

12.3.1 对应分析的基本原理

对应分析的实质是要寻找一个能建立起 R 型和 Q 型分析内在联系的一个过渡矩阵 Z ,使得变量点的协差阵 $A=Z'Z$ 和样品点的协差阵 $B=ZZ'$ 。这样两者具有相同的非零特征根 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$, 其中 $0 < m \leq \min(p, n)$ 。假设 A 的特征根 λ_i 对应的特征向量为 U_i , 则 B 的特征根 λ_i 对应的特征向量就是 $Z_{ui}=V_i$ 。这样借助于 R 型因子分析的结果就能很容易得到 Q 型因子分析的结果,从而实现一步双解的目的。

过渡矩阵 Z 的构建,这可以通过以下的方法来实现。

设有 n 个样品,每个样品有 p 个指标,记 x_{ij} 为第 i 个样品在第 j 个指标上的测试值,且设 $x_{ij} > 0$,如果实测值有小于 0 的情况出现时,只需对所有观测值都加上一个足够大的正数,便能满足这一要求。由此可得到如表 12-15 所示的 np 个原始观测数据。

表中, $x_{i.}$ 表示第 i 行的和 ($x_{i.} = \sum_{j=1}^p x_{ij}$), $x_{.j}$ 表示第 j 列的列和 ($x_{.j} = \sum_{i=1}^n x_{ij}$), T 表示

所有观测值的总和 ($T = \sum_{i=1}^n \sum_{j=1}^p x_{ij}$)。

表 12-15 原始观测数据记录表

指 标 样 品	指标 1	指标 2	...	指标 p	总和
1	x_{11}	x_{12}	...	x_{1p}	$x_{1.}$
2	x_{21}	x_{22}	...	x_{2p}	$x_{2.}$
\vdots	\vdots	\vdots		\vdots	\vdots
n	x_{n1}	x_{n2}	...	x_{np}	$x_{n.}$
总和	$x_{.1}$	$x_{.2}$...	$x_{.p}$	T

对上表中的所有原始观测数据和其边际和都除以总和 T ，这相当于改变了测度尺度，使变量和样品具有相同比例大小，即 $p_{ij}=x_{ij}/T$ ，显然 $0 < p_{ij} < 1$ ，并且 $\sum_{i=1}^n \sum_{j=1}^p p_{ij} = 1$ ，所以， p_{ij} 可解释为“概率”，这样可将原始数据转换成规格化的“概率”，得到表 12-16。

表 12-16 转换结果记录表 ($p_{ij}=x_{ij}/T$)

指 标 样 品	指标 1	指标 2	...	指标 p	总和
1	p_{11}	p_{12}	...	p_{1p}	$p_{1.}$
2	p_{21}	p_{22}	...	p_{2p}	$p_{2.}$
\vdots	\vdots	\vdots		\vdots	\vdots
n	p_{n1}	p_{n2}	...	p_{np}	$p_{n.}$
总和	$p_{.1}$	$p_{.2}$...	$p_{.p}$	1

假如，将 n 个样品看成 p 维空间的点，那么 n 个点中的第 i 个点的坐标可用 $\left(\frac{p_{i1}}{p_{.1}}, \frac{p_{i2}}{p_{.2}}, \dots, \frac{p_{ip}}{p_{.p}} \right)$ 来表示，其中 $j=1, 2, \dots, p$ ，称它为 n 个样品点。

如果引入欧氏距离来刻划两个样品间的相对关系，则此时任意两个样品点 K 和 L 之间的加权欧氏距离为

$$D^2(K, L) = \sum_{j=1}^p \left(\frac{p_{Kj}}{p_{.j}} - \frac{p_{Lj}}{p_{.j}} \right)^2 \cdot \frac{1}{p_{.j}} = \sum_{j=1}^p \left(\frac{p_{Kj}}{\sqrt{p_{.j} p_{.j}}} - \frac{p_{Lj}}{\sqrt{p_{.j} p_{.j}}} \right)^2$$

这相当于是 n 个点的坐标为

$$\left(\frac{p_{i1}}{\sqrt{p_{.1} p_{.1}}}, \frac{p_{i2}}{\sqrt{p_{.2} p_{.2}}}, \dots, \frac{p_{ip}}{\sqrt{p_{.p} p_{.p}}} \right) \quad j=1, 2, \dots, p$$

中任意两个样品点 K 和 L 之间的欧氏距离。

同样地, 将 p 个变量看成是 n 维空间的点, 则 p 个变量点为 $\left(\frac{p_{1j}}{p_{.j}}, \frac{p_{2j}}{p_{.j}}, \dots, \frac{p_{nj}}{p_{.j}}\right)$, $j=1, 2, \dots, p$ 。此时两个变量 i 和 j 之间的加权欧氏距离为

$$D^2(i, j) = \sum_{k=1}^n \left(\frac{p_{ki}}{\sqrt{p_{k.} p_{.i}}} - \frac{p_{kj}}{\sqrt{p_{k.} p_{.j}}} \right)^2$$

为给出变量点协差阵和样品点协差阵的定义, 首先要给出计算样品点中第 i 个变量的均值的计算公式, 由加权算术平均数的定义可知, 样品点中第 i 个变量的均值的计算公式为

$$\sum_{i=1}^n \frac{p_{ij}}{\sqrt{p_{.j} p_{.i}}} p_{.i} = \frac{\sum_{i=1}^n p_{ij}}{\sqrt{p_{.j}}} = \frac{p_{.j}}{\sqrt{p_{.j}}} = \sqrt{p_{.j}} \quad j=1, 2, \dots, p$$

可以验证, 它不仅是所有样品平均点坐标, 而且还是各变量的平均值。故样品空间中变量点的协差阵为

$$A=(a_{ij})$$

其中

$$\begin{aligned} a_{ij} &= \sum_{\alpha=1}^n \left(\frac{p_{\alpha i}}{\sqrt{p_{.i} p_{\alpha.}}} - \sqrt{p_{.i}} \right) \left(\frac{p_{\alpha j}}{\sqrt{p_{.j} p_{\alpha.}}} - \sqrt{p_{.j}} \right) p_{\alpha.} \\ &= \sum_{\alpha=1}^n \left(\frac{p_{\alpha i} - p_{.i} p_{\alpha.}}{\sqrt{p_{.i} p_{\alpha.}}} \right) \left(\frac{p_{\alpha j} - p_{.j} p_{\alpha.}}{\sqrt{p_{.j} p_{\alpha.}}} \right) \\ &= \sum_{\alpha=1}^n z_{\alpha i} z_{\alpha j} \end{aligned}$$

其中

$$z_{\alpha i} = \frac{p_{\alpha i} - p_{.i} p_{\alpha.}}{\sqrt{p_{.i} p_{\alpha.}}} = \frac{\frac{x_{\alpha i}}{T} - \frac{x_{.i}}{T} \cdot \frac{x_{\alpha.}}{T}}{\sqrt{\frac{x_{.i}}{T} \cdot \frac{x_{\alpha.}}{T}}} = \frac{x_{\alpha i} - \frac{x_{.i} x_{\alpha.}}{T}}{\sqrt{x_{.i} x_{\alpha.}}}$$

$$\alpha = 1, 2, \dots, n \quad i = 1, 2, \dots, p$$

令 $Z=(z_{ij})$, 则有

$$A=Z' Z$$

同理可得, 样品点的协差阵 $B=(b_{KL})$

其中

$$\begin{aligned}
 b_{KL} &= \sum_{i=1}^p \left(\frac{p_{Ki}}{\sqrt{p_{K.} p_{i.}}} - \sqrt{p_{K.}} \right) \left(\frac{p_{Li}}{\sqrt{p_{L.} p_{i.}}} - \sqrt{p_{L.}} \right) p_{i.} \\
 &= \sum_{i=1}^p \left(\frac{p_{Ki} - p_{i.} p_{K.}}{\sqrt{p_{i.} p_{K.}}} \right) \left(\frac{p_{Li} - p_{i.} p_{L.}}{\sqrt{p_{i.} p_{L.}}} \right) \\
 &= \sum_{i=1}^p z_{Ki} z_{Li}
 \end{aligned}$$

其中

$$\begin{aligned}
 z_{Ki} &= \frac{p_{Ki} - p_{i.} p_{K.}}{\sqrt{p_{i.} p_{K.}}} = \frac{x_{Ki} - \frac{x_{i.} x_{K.}}{T}}{\sqrt{x_{i.} x_{K.}}} \\
 z_{Li} &= \frac{p_{Li} - p_{i.} p_{L.}}{\sqrt{p_{i.} p_{L.}}} = \frac{x_{Li} - \frac{x_{i.} x_{L.}}{T}}{\sqrt{x_{i.} x_{L.}}}
 \end{aligned}$$

因此,

$$B = ZZ'$$

由上可知, 当将原始数据阵

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

变换成 Z 时, 则变量点的协差阵为 $A = Z'Z$, 而样品点的协差阵为 $B = ZZ'$, 两者间明显存在简单的对应关系, 并且将原始数据 x_{ij} 变成 z_{ij} 后, z_{ij} 对于 i, j 是对称的, 也就是 z_{ij} 对变量和样品具有对等性。

根据线性代数中的相关定理可知, A 与 B 的非零特征根相同。并且, 如果 U 是 $Z'Z$ 的特征向量, 则 ZU 是 ZZ' 的特征向量。如果 V 是 ZZ' 的特征向量, 则 $Z'V$ 是 $Z'Z$ 的特征向量。由此建立了因子分析中 R 型和 Q 型的关系。因此, 从 R 型因子分析出发可以直接获得 Q 型因子分析的结果, 另外, 它还可以把变量点和样品点同时反映在同一个因子轴所确定的平面上, 从而根据接近程度, 将变量点和样品点一起考虑进行分类。

虽然对应分析和因子分析一样都是一种在比变量的最小类别数还小的低维度空间中描述变量间关系的一种实用多元统计分析技术。但两者间还是有区别的, 对于因子分析而言, 它要求等间隔数据, 而且观测量数目必须是变量数的 5 倍, 它只能分别地对指标或样品进行分类。相反, 对应分析假定行、列变量为名义变量, 不但可以很好地描述每个变量类别间的关系, 还可以描述这些变量之间的关系。另外, 对应分析可用于对任何正的对应对应的表格进行分析。

12.3.2 对应分析实例分析

1. 变量为等间隔测度

例 12.3 仍以例 12.2 中的数据资料为例, 试对全国 30 个省、市、自治区与其经济发展基本情况做对应分析。数据存放在 data12-03.sav 中。

重要提示:

虽然, 本例中所用到的数据资料同前面例 12.2 因子分析中所用到的数据资料是同一批数据资料, 但是, 需要记住的是对应分析中所用到的数据文件同因子分析中所用到的数据文件的格式是不一样的, 所以, 本例用的数据文件为 data12-03.sav, 而不是 data12-02.sav。

关于对应分析中所要建立的数据文件的一些说明。

对应分析中, 未做任何处理的原始数据是无效的, 有效的数据必须是要事先制作成同交互式列表中的单元格计数资料类型一样的资料, 在对应分析前先用 WEIGHT 命令来进行处理。因此, 在对应分析的数据文件中, 须定义三个变量。将要放在对应分析过程的行和列中的变量是分类变量, 需用名义测度定义。第三个变量是对应行、列的实际测试值, 一般为尺度变量。

所建的数据文件, 在被对应分析程序调用前, 应先到 Data 的下拉式菜单 Weight Cases 的对话框中, 选择非默认选择项 Weight cases by, 将第三个变量选中, 单击右箭头按钮键, 将其移入 Frequency Variable 的下框中, 做加权处理。权重即是第三个变量的数值。如果表格的任意一个单元格中有 0, 即第三个变量中有记录值 0, 则 WEIGHT 命令处理过程中会发出警告提示, 但它不影响对应分析的正常分析工作, 经过这一步处理后, 可以对其做对应分析。

在本例 data12-03.sav 的数据文件中, 共有三个变量, 分别为 *省份*、*指标*和*观测值*。*省份*表示样品分类指标, 名义变量, 从 1 到 30 共分 30 类, 分别代表 30 个不同的省、市、自治区, 具体数字所代表的省、市、自治区参见省份变量值标签中的内容; *指标*表示反映经济发展基本情况的八项实测指标, 名义变量, 从 1 到 8 共分 8 类, 具体数字所代表的具体测试指标参见指标变量值标签中的内容; *观测值*表示这些省份对应反映经济发展基本情况的八项指标的观测值, 尺度变量。

在 SPSS 中, 进行对应分析的步骤如下:

(1) 打开数据文件

在数据编辑窗口中, 打开 data12-03.sav。并按上面的要求对数据文件用观测值作为权重变量进行加权处理。

(2) 定义行、列变量

按 Analyze → Data Reduction → Correspondence Analysis 顺序, 展开 Correspondence

Analysis 对话框, 见图 12-8。

选择省份, 单击右箭头按钮将其移入 Row 的下框中, 单击其下的 Define Ranges 按钮, 展开如图 12-9 所示的 Correspondence Analysis: Define Row 对话框, 在 Minimum value 中输入 1, 在 Maximum value 中输入 30; 单击 Update 按钮, 将省份定义的 1 到 30 代表 30 个不同省份的种类值上传到 Category Constraints 的下框中。由于没有增补项、等同约束项及强制性等同约束项, 因此, 在 Category Constraints 的选择项中使用系统默认选择项 None。

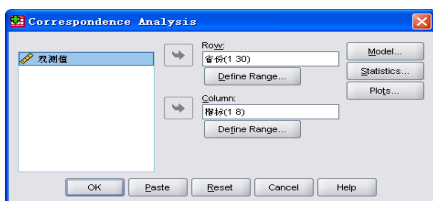


图 12-8 Correspondence Analysis 对话框

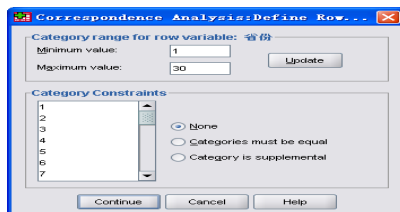


图 12-9 Correspondence Analysis: Define Row 对话框

单击 Continue 按钮, 返回 Correspondence Analysis 对话框 (见图 12-8)。

选择指标变量, 按右箭头按钮将其移入 Column 的下框中, 单击其下的 Define Ranges 按钮, 展开同图 12-9 几乎完全一样的 Correspondence Analysis: Define Column 对话框。在 Minimum value: 中输入 1, 在 Maximum value 中输入 8。

单击 Continue 按钮, 返回 Correspondence Analysis 对话框 (见图 12-8)。

(3) 定义解的维度、距离测度、标准化方法和正规化方法

单击 Model 按钮, 展开 Correspondence Analysis: Model 对话框, 在 Dimensions in solution: (解的维度) 选择项中, 输入 2, 即将样品和指标在二维空间中对应地进行分类。

在 Distance Measure 距离测度上, 选择系统默认值 Chi square 距离测度 (标准对应分析要求使用本测度)。

在 Standardization Method 标准化方法选择项上, 由于在上面的距离测度选择项中选中了卡方距离测度, 故在此只能选择系统默认选择项 “Row and column means are removed”, 即同时对行和列两者进行中心化处理。

在 Normalization Method 正规化方法选择项中, 选择 Symmetrical, 即使用对称法。使用本法可以检查两个变量的类别间的差异或相似, 对各个维度而言, 行得分是列得分除以匹配奇异值的加权平均, 列得分是行得分除以匹配奇异值的加权平均。

单击 Continue 按钮, 返回 Correspondence Analysis 对话框 (见图 12-8)。

(4) 指定输出结果表的种类

单击 Statistics 按钮, 弹出 Correspondence Analysis: Statistics 对话框, 见图 12-11。在

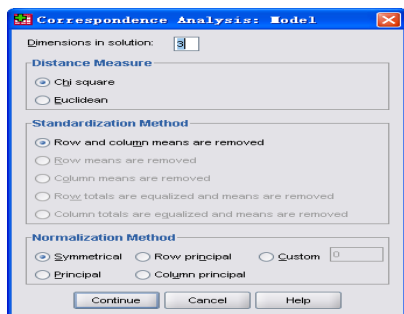


图 12-10 Correspondence Analysis:
Model 对话框

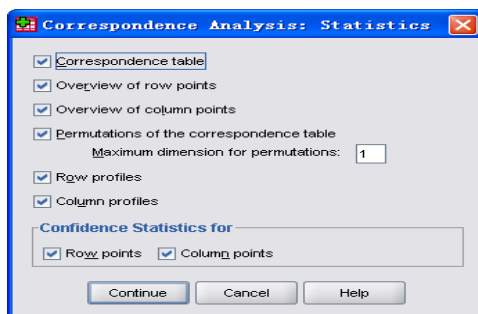


图 12-11 Correspondence Analysis:
Statistics 对话框

本对话框中，选择 Correspondence table（对应表）选择项，要求在输出窗中输出含有行和列的边际总和在内的输入变量的交叉分组列表；选择 Overview of row points（行分数综述）选择项，要求在输出窗中为各个行类别显示包括得分、质量、惯量、对维度的惯量贡献、维度对分数惯量的贡献的综合表；选择 Overview of column points（列分数综述）选择项，要求在输出窗中为各个列类别显示包括得分、质量、惯量、对维度的惯量贡献、维度对分数惯量的贡献的综合表；选择 Row profiles 选择项，要求在输出窗中为各个行类别显示包括列变量类别的横向分布；选择 Column profiles 选择项，要求在输出窗中为各个列类别显示包括行变量类别的横向分布。

在 Confidence Statistics for 选择项中，选择 Row points 选择项，要求在输出窗中显示包括标准差和所有非增补行分数相关内容在内的表格；选择 Column points 选择项，要求在输出窗中显示包括标准差和所有非增补列分数相关内容在内的表格。

单击 Continue 按钮，返回 Correspondence Analysis 对话框（见图 12-8）。

（5）选定产生的图形

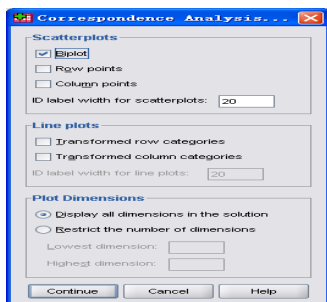


图 12-12 Correspondence
Analysis:Plots 对话框

单击 Plots 按钮，弹出如图 12-12 所示的 Correspondence Analysis:Plots 对话框。在 Scatterplots（散点图）选择项中，选择 Biplot（双维图法）选择项。要求在输出窗中产生矩阵的行、列分数的联合图。需要注意的是，如果在正规化方法中选择了 Principal 选择项，则，本选择项无效。

其他保持系统默认选项。

单击 Continue 按钮，返回 Correspondence Analysis 对话框（见图 12-8）。

（6）结果与讨论

单击 OK 按钮，执行运算。则在输出窗口中得到所

要求的输出表，具体见表 12-17 至表 12-24，以及输出图，见图 12-13。

表 12-17 对应表

省份	Correspondence Table								
	GDP	居民消费水平	固定资产投资	职工平均工资	货物周转量	居民消费价格指数	商品零售价格指数	工业总产值	Active Margin
北京	1394.890	2505.000	519.010	8144.000	373.900	117.300	112.600	843.430	14010.130
天津	920.110	2720.000	345.460	6501.000	342.800	115.200	110.600	582.510	11637.680
河北	2849.520	1258.000	704.870	4839.000	2033.300	115.200	115.800	1234.850	13150.540
山西	1092.480	1250.000	290.900	4721.000	717.300	116.900	115.600	697.250	9001.430
内蒙	832.880	1387.000	250.230	4134.000	781.700	117.500	116.800	419.390	8039.500
辽宁	2793.370	2397.000	387.990	4911.000	1371.100	116.100	114.000	1840.550	13931.110
吉林	1129.200	1872.000	320.450	4430.000	497.400	115.200	114.200	726.470	9204.920
黑龙江	2014.530	2343.000	435.730	4145.000	824.800	116.100	114.300	1240.370	11233.830
上海	2462.570	5343.000	996.480	9279.000	207.400	118.700	113.000	1642.950	20163.100
江苏	5155.250	1926.000	1434.950	5943.000	1025.500	115.800	114.300	2026.640	17741.440
浙江	3524.790	2249.000	1006.390	6619.000	754.400	116.600	113.500	916.590	15300.270
安徽	2003.580	1254.000	474.000	4609.000	908.300	114.870	112.700	824.140	10300.590
福建	2160.520	2320.000	553.970	5857.000	609.300	115.200	114.400	433.670	12164.060
江西	1205.110	1182.000	282.840	4211.000	411.700	116.900	115.900	571.840	8097.290
山东	5002.340	1527.000	1229.550	5145.000	1196.600	117.600	114.200	2207.690	16539.980
河南	3002.740	1034.000	670.350	4344.000	1574.400	116.500	114.900	1367.920	12224.810
湖北	2391.420	1527.000	571.680	4685.000	849.000	120.000	116.600	1220.720	11481.420
湖南	2195.700	1408.000	422.610	4797.000	1011.800	119.000	115.500	843.830	10913.440
广东	5381.720	2699.000	1639.830	8250.000	656.500	114.000	111.600	1396.350	20249.000
广西	1606.150	1314.000	382.590	5105.000	556.000	118.400	116.400	554.970	9753.510
海南	364.170	1814.000	198.350	5340.000	232.100	113.500	111.300	64.330	8237.750
四川	3534.000	1261.000	822.540	4645.000	902.300	118.500	117.000	1431.810	12832.150
贵州	630.070	942.000	150.840	4475.000	301.100	121.400	117.200	324.720	7062.330
云南	1206.680	1261.000	334.000	5149.000	310.400	121.300	118.100	716.650	9217.130
西藏	55.980	1110.000	17.870	7382.000	4.200	117.300	114.900	5.570	8807.820
陕西	1000.030	1208.000	300.270	4396.000	500.900	119.000	117.000	600.980	8242.180
甘肃	553.350	1007.000	114.810	5493.000	507.000	119.800	116.500	468.790	8380.250
青海	165.310	1445.000	47.760	5753.000	61.600	118.000	116.300	105.800	7812.770
宁夏	169.750	1355.000	61.980	5079.000	121.800	117.100	115.300	114.400	7134.330
新疆	834.570	1469.000	376.950	5348.000	339.000	119.700	116.700	428.760	9032.680
Active Margin	57633.780	52387.000	15345.250	163729.000	19983.600	3518.670	3447.200	25853.940	341897.440

表 12-17 显示了用来进行分析的原始数据资料的记录表，另外它还包括行、列的边际和及总和。用它可以同表 12-6 中的原始数据进行核对，以此来确定输入数据资料的正确性。

表 12-18 行归一化处理表

省份	Row Profiles								
	GDP	居民消费水平	固定资产投资	职工平均工资	货物周转量	居民消费价格指数	商品零售价格指数	工业总产值	Active Margin
北京	.100	.179	.037	.581	.027	.008	.008	.060	1.000
天津	.079	.234	.030	.559	.029	.010	.010	.050	1.000
河北	.217	.096	.054	.368	.155	.009	.009	.094	1.000
山西	.121	.139	.032	.524	.080	.013	.013	.077	1.000
内蒙	.104	.173	.031	.514	.097	.015	.015	.052	1.000
辽宁	.201	.172	.028	.353	.098	.008	.008	.132	1.000
吉林	.123	.203	.035	.481	.054	.013	.012	.079	1.000
黑龙江	.179	.209	.039	.369	.073	.010	.010	.110	1.000
上海	.122	.265	.049	.460	.010	.006	.006	.081	1.000
江苏	.291	.109	.081	.335	.058	.007	.006	.114	1.000
浙江	.230	.147	.066	.433	.049	.008	.007	.060	1.000
安徽	.195	.122	.046	.447	.088	.011	.011	.080	1.000
福建	.178	.191	.046	.482	.050	.009	.009	.036	1.000
江西	.149	.146	.035	.520	.051	.014	.014	.071	1.000
山东	.302	.092	.074	.311	.072	.007	.007	.133	1.000
河南	.246	.085	.055	.355	.129	.010	.009	.112	1.000
湖北	.208	.133	.050	.408	.074	.010	.010	.106	1.000
湖南	.201	.129	.039	.440	.093	.011	.011	.077	1.000
广东	.266	.133	.081	.407	.032	.006	.006	.069	1.000
广西	.165	.135	.039	.523	.057	.012	.012	.057	1.000
海南	.044	.220	.024	.648	.028	.014	.014	.008	1.000
四川	.275	.098	.064	.362	.070	.009	.009	.112	1.000
贵州	.089	.133	.021	.634	.043	.017	.017	.046	1.000
云南	.131	.137	.036	.559	.034	.013	.013	.078	1.000
西藏	.006	.126	.002	.838	.000	.013	.013	.001	1.000
陕西	.121	.147	.036	.533	.061	.014	.014	.073	1.000
甘肃	.066	.120	.014	.655	.060	.014	.014	.056	1.000
青海	.021	.185	.006	.736	.008	.015	.015	.014	1.000
宁夏	.024	.190	.009	.712	.017	.016	.016	.016	1.000
新疆	.092	.163	.042	.592	.038	.013	.013	.047	1.000
Mass	.169	.153	.045	.479	.058	.010	.010	.078	1.000

表 12-18 显示了用行边际和作为分母逐行进行归一化处理的结果。最后一行的 Mass

为列的边际“概率”，也即各列和占总和中的百分比。

表 12-19 列归一化处理表

省份	Column Profiles								
	GDP	居民消费水平	固定资产投资	职工平均工资	货物周转量	居民消费价格指数	商品零售价格指数	工业增加值	Mass
北京	.024	.048	.034	.050	.019	.033	.033	.033	.041
天津	.016	.052	.023	.040	.017	.033	.032	.023	.034
河北	.049	.024	.046	.030	.102	.033	.034	.048	.038
山西	.019	.024	.019	.029	.036	.033	.034	.027	.026
内蒙	.014	.026	.016	.025	.039	.033	.034	.016	.024
辽宁	.048	.046	.025	.030	.069	.033	.033	.071	.041
吉林	.020	.036	.021	.027	.025	.033	.033	.028	.027
黑龙江	.035	.045	.028	.025	.041	.033	.033	.048	.033
上海	.043	.102	.065	.057	.010	.034	.033	.064	.059
江苏	.089	.037	.094	.036	.051	.033	.033	.078	.052
浙江	.061	.043	.066	.040	.038	.033	.033	.035	.045
安徽	.035	.024	.031	.028	.045	.033	.033	.032	.030
福建	.037	.044	.036	.036	.030	.033	.033	.017	.036
江西	.021	.023	.018	.026	.021	.033	.034	.022	.024
山东	.087	.029	.080	.031	.060	.033	.033	.085	.048
河南	.052	.020	.044	.027	.079	.033	.033	.053	.036
湖北	.041	.029	.037	.029	.042	.034	.034	.047	.034
湖南	.038	.027	.028	.029	.051	.034	.034	.033	.032
广东	.093	.052	.107	.050	.033	.032	.032	.054	.059
广西	.028	.025	.025	.031	.028	.034	.034	.021	.029
海南	.006	.035	.013	.033	.012	.032	.032	.002	.024
四川	.061	.024	.054	.028	.045	.034	.034	.055	.038
贵州	.011	.018	.010	.027	.015	.035	.034	.013	.021
云南	.021	.024	.022	.031	.016	.034	.034	.028	.027
西藏	.001	.021	.001	.045	.000	.033	.033	.000	.026
陕西	.017	.023	.020	.027	.025	.034	.034	.023	.024
甘肃	.010	.019	.007	.034	.025	.034	.034	.018	.025
青海	.003	.028	.003	.035	.003	.034	.034	.004	.023
宁夏	.003	.026	.004	.031	.006	.033	.033	.004	.021
新疆	.014	.028	.025	.033	.017	.034	.034	.017	.026
Active Margin	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	

表 12-19 是同表 12-18 很相似的一张表，它显示了列边际和作为分母逐列进行归一化处理的结果。最后一列的 Mass 为行的边际“概率”，也即各行和占总和中的百分比。

表 12-20 汇总表

Dimension	Summary							
	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation
1	.314	.099			.740	.740	.001	
2	.134	.018			.136	.876	.002	
3	.108	.012			.088	.964		
4	.060	.004			.027	.991		
5	.027	.001			.006	.997		
6	.021	.000			.003	1.000		
7	.001	.000			.000	1.000		
Total		.133	45578.350	.000*	1.000	1.000		.029

a. 203 degrees of freedom

表 12-20 从左向右各列依次为维度数、奇异值（惯量的平方根）、惯量（就是常说的特征根，它用来说明对应分析各个维度的结果能够解释列联表中两个变量联系的程度）、卡方值、P 值、惯量的比例（占总方差的百分比、累计百分比）、置信奇异值（前二个维度的标准差、前二个维度相互间的相关系数）。从该表可见，第一个维度的惯量值为 0.099、奇异值为 0.314，第二个维度的惯量值为 0.018、奇异值为 0.134，它们分别解释了总信息量的 74% 和 13.6% 的信息，前两个维度累计解释了总信息量的 87.6%。因此，二维图形基本已可以表示两个变量间的信息。

表 12-21 左侧第一列为变量省份的 30 个取值类别，第二列为其质量（Mass）值，也就是样品的每个类别所占的百分比，同表 12-19 最右侧列的结果完全一样。第三列为每个

样品在两个维度中的分值，也就是在平面直角坐标系中的坐标值。第四列为每个样品的惯量值。第五列为每个样品对各个维度的贡献量，包括样品点对维度惯量的贡献和维度对样品点惯量的贡献两种。

表 12-21 样品点坐标值

Overview Row Points ^a									
省份	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension	Of Dimension to Inertia of Point	Of Point to Inertia of Dimension	Of Dimension to Inertia of Point	Total
北京	.041	-.470	-.162	.003	.029	.008	.903	.046	.949
天津	.034	-.583	-.270	.004	.037	.018	.813	.075	.888
河北	.038	.617	.807	.009	.047	.186	.530	.388	.918
山西	.026	-.157	.397	.001	.002	.031	.246	.671	.917
内蒙	.024	-.247	.467	.002	.005	.038	.269	.413	.682
辽宁	.041	.425	.186	.005	.023	.010	.479	.039	.519
吉林	.027	-.205	-.082	.001	.004	.001	.403	.028	.431
黑龙江	.033	.225	-.133	.002	.005	.004	.238	.036	.274
上海	.059	-.283	-.778	.008	.015	.265	.181	.583	.764
江苏	.052	.741	-.306	.010	.091	.036	.881	.064	.945
浙江	.045	.232	-.280	.002	.008	.026	.391	.244	.634
安徽	.030	.221	.297	.001	.005	.020	.546	.423	.969
福建	.036	-.121	-.214	.001	.002	.012	.142	.192	.334
江西	.024	-.162	.089	.000	.002	.001	.625	.080	.705
山东	.048	.882	-.135	.012	.120	.007	.946	.009	.956
河南	.036	.707	.576	.007	.057	.088	.776	.220	.996
湖北	.034	.338	.070	.001	.012	.001	.906	.017	.922
湖南	.032	.232	.322	.001	.005	.025	.491	.406	.896
广东	.059	.395	-.520	.007	.029	.119	.426	.316	.743
广西	.029	-.115	.125	.000	.001	.003	.333	.167	.500
海南	.024	-.905	-.068	.006	.063	.001	.956	.002	.958
四川	.038	.653	-.039	.005	.051	.000	.931	.001	.932
贵州	.021	-.575	.303	.003	.022	.014	.818	.097	.916
云南	.027	-.271	.014	.001	.006	.000	.621	.001	.622
西藏	.026	-1.303	.294	.016	.139	.017	.881	.019	.901
陕西	.024	-.224	.213	.001	.004	.008	.634	.245	.879
甘肃	.025	-.603	.557	.004	.028	.057	.694	.254	.948
青海	.023	-1.130	.058	.009	.093	.001	.985	.001	.987
宁夏	.021	-1.065	.096	.007	.075	.001	.991	.003	.994
新疆	.026	-.493	.019	.002	.020	.000	.937	.001	.937
Active Total	1.000			.133	1.000	1.000			

a. Symmetrical normalization

表 12-22 为变量指标的八个类别在两个维度中的分值，具体含义同表 12-21 中的解释，所不同的是本表是说明变量点的。

表 12-22 变量点坐标值

Overview Column Points ^a									
指标	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension	Of Dimension to Inertia of Point	Of Point to Inertia of Dimension	Of Dimension to Inertia of Point	Total
GDP	.169	.839	-.190	.040	.378	.045	.940	.021	.960
居民消费水平	.153	-.311	-.450	.015	.047	.230	.317	.284	.602
固定资产投资	.045	.675	-.529	.009	.065	.093	.676	.178	.854
职工平均工资	.479	-.440	.101	.031	.295	.037	.927	.021	.948
货物周转量	.058	.700	1.134	.021	.091	.559	.427	.480	.907
居民消费价格指数	.010	-.349	.479	.001	.004	.018	.412	.332	.745
商品零售价格指数	.010	-.345	.493	.001	.004	.018	.396	.345	.742
工业总产值	.076	.694	-.001	.015	.116	.000	.756	.000	.756
Active Total	1.000			.133	1.000	1.000			

a. Symmetrical normalization

表 12-23 给出了各个样品在两个维度上的标准差值和两个维度上的相关系数值。

表 12-24 给出量各个变量在两个维度上的标准差值和两个维度上的相关系数值。

从图 12-13 可见，在 1996 年，各地经济发展速度、发展特点、消费水平已有了明显的不同，地域和经济之间的联系已有了初步的特征，可将变量点和样品点分为五类：

表 12-23 置信样品点

Confidence Row Points			
省份	Standard Deviation in Dimension		Correlation
	1	2	1-2
北京	.007	.011	-.089
天津	.007	.021	-.076
河北	.014	.024	.034
山西	.007	.011	.079
内蒙	.010	.024	.045
辽宁	.012	.040	.021
吉林	.008	.023	-.006
黑龙江	.009	.036	-.017
上海	.011	.030	-.054
江苏	.007	.018	.077
浙江	.007	.018	.008
安徽	.004	.005	-.176
福建	.010	.017	-.041
江西	.008	.013	.035
山东	.006	.016	.107
河南	.007	.008	-.304
湖北	.005	.009	-.040
湖南	.009	.014	-.028
广东	.009	.027	.002
广西	.004	.013	.009
海南	.007	.013	-.026
四川	.006	.017	.057
贵州	.012	.025	.069
云南	.007	.017	.044
西藏	.011	.039	-.017
陕西	.007	.010	.033
甘肃	.008	.014	.162
青海	.006	.015	-.014
宁夏	.005	.010	.070
新疆	.008	.014	.064

第一类:

变量点为固定资产投资、GDP、工业总产值。

样品点为江苏、山东、广东和四川。

它们是 20 世纪 90 年代中后期国家重点发展地区,具有高投入、高产出的经济结构特征。

第二类:

在投入、产出方面与第一类正相反,宁夏、青海、西藏属低投入、低产出的经济欠发达地区。

第三类:

变量点为货物周转量。

样品点为河北、河南。

由于发展布局的不同,它们已成为当时全国货物的集散地。

第四类:

变量点为居民消费水平。

样品点为上海。

发展较早的上海已步入当时的高消费时代。

第五类:

处在中间的 20 个省、市、自治区有湖南、湖北、安徽、黑龙江、浙江、辽宁、天津、福建、北京、吉林、海南、云南、新疆、江西、广西、陕西、贵州、山西、内蒙、甘肃。

它们在固定资产投入、产出、工资水平、消费水平等方面有相似的地方,都处于中等平均水平,是我国当时经济发展整体水平的代表。

表 12-24 置信变量点

Confidence Column Points			
指标	Standard Deviation in Dimension		Correlation
	1	2	1-2
GDP	.005	.017	.108
居民消费水平	.008	.032	-.033
固定资产投资	.014	.030	.026
职工平均工资	.003	.010	.104
货物周转量	.014	.031	-.037
居民消费价格指数	.031	.043	-.045
商品零售价格指数	.032	.043	-.044
工业总产值	.010	.029	.045

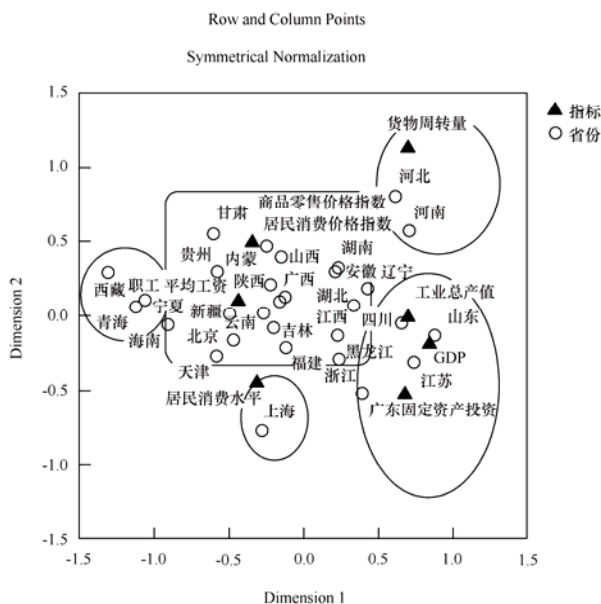


图 12-13 对应分析图

2. 变量为名义测度

例 12.4 在对 218 名受访人员进行收入水平和品牌选择关系的调查研究中，得到表 12-25 中的调查数据，数据已存放在 data12-04.sav 中，试对其进行对应分析。

表 12-25 收入水平与品牌选择

		收入水平		
		低	中	高
品牌	A	2	7	16
	B	49	7	3
	C	4	5	23
	D	4	49	5
	E	15	2	5
	F	1	7	14

(1) 打开数据文件

在数据编辑窗口中，打开 data12-04.sav。并按上面的要求对数据文件用 *观测值* 作为权重变量进行加权处理。

(2) 定义行、列变量

按 Analyze → Data Reduction → Correspondence Analysis 顺序，展开 Correspondence Analysis 对话框，见图 12-8。

选择 *品牌*，单击右箭头按钮将其移入 Row 的下框中，单击其下的 Define Ranges 按钮，

展开如图 12-9 所示的 Correspondence Analysis:Define Row 对话框, 在 Minimum value:中输入 1, 在 Maximum value:中输入 6; 单击 Update 按钮, 将品牌定义的 1 到 6 代表 6 个不同品牌的类别值上传到 Category Constraints 的下框中。由于没有增补项、等同约束项及强制性等同约束项, 因此, 在 Category Constraints 的选择项中使用系统默认选择项 None。

单击 Continue 按钮, 返回 Correspondence Analysis 对话框 (见图 12-8)。

选择收入水平变量, 单击右箭头按钮将其移入 Column 的下框中, 单击其下的 Define Ranges 按钮, 展开同图 12-9 几乎完全一样的 Correspondence Analysis:Define Column 对话框。在 Minimum value:中输入 1, 在 Maximum value:中输入 3。

单击 Continue 按钮, 返回 Correspondence Analysis 对话框 (见图 12-8)。

(3) 定义解的维度、距离测度、标准化方法和正规化方法

单击 Model 按钮, 展开 Correspondence Analysis:Model 对话框, 在 Dimensions in solution: (解的维度) 选择项中, 输入 2, 即将行、列变量在二维空间中对应地进行分类。

在 Distance Measure 距离测度上, 选择系统默认值 Chi square 距离测度 (标准对应分析要求使用本测度)。

在 Standardization Method 标准化方法选择项上, 由于在上面的距离测度选择项中选定了卡方距离测度, 故在此只能选择系统默认选择项“Row and column means are removed”, 即同时对行和列两者进行中心化处理。

在 Normalization Method 正规化方法选择项中, 选择 Symmetrical, 即使用对称法。使用本法可以检查两个变量的类别间的差异或相似, 对各个维度而言, 行得分是列得分除以匹配奇异值的加权平均, 列得分是行得分除以匹配奇异值的加权平均。

单击 Continue 按钮, 返回 Correspondence Analysis 对话框 (见图 12-8)。

(4) 指定输出结果表的种类

单击 Statistics 按钮, 弹出 Correspondence Analysis:Statistics 对话框, 见图 12-11。在本对话框中, 选择 Correspondence table (对应表) 选择项, 要求在输出窗中输出含有行和列的边际总和在内的输入变量的交叉分组列表; 选择 Overview of row points (行分数综述) 选择项, 要求在输出窗中为各个行类别显示包括得分、质量、惯量、对维度的惯量贡献、维度对分数惯量的贡献的综合表; 选择 Overview of column points (列分数综述) 选择项, 要求在输出窗中为各个列类别显示包括得分、质量、惯量、对维度的惯量贡献、维度对分数惯量的贡献的综合表; 选择 Row profiles 选择项, 要求在输出窗中为各个行类别显示包括列变量类别的横向分布; 选择 Column profiles 选择项, 要求在输出窗中为各个列类别显示包括行变量类别的横向分布。

在 Confidence Statistics for 选择项中, 选择 Row points 选择项, 要求在输出窗中显示包括标准差和所有非增补行分数相关内容在内的表格; 选择 Column points 选择项, 要求在输出窗中显示包括标准差和所有非增补列分数相关内容在内的表格。

单击 Continue 按钮, 返回 Correspondence Analysis 对话框 (见图 12-8)。

(5) 选定产生的图形

单击 Plots 按钮, 弹出如图 12-12 所示的 Correspondence Analysis:Plots 对话框。在 Scatterplots (散点图) 选择项中, 选择 Biplot (双维图法) 选择项。要求在输出窗中产生矩阵的行、列分数的联合图。需要注意的是, 如果在正规化方法中选择了 Principal 选择项, 则, 本选择项无效。

其他保持系统默认选项。

单击 Continue 按钮, 返回 Correspondence Analysis 对话框 (见图 12-8)。

(6) 结果与讨论

单击 OK 按钮, 执行运算。则在输出窗口中得到所要求的输出表, 具体见表 12-26 至表 12-29, 以及输出图, 见图 12-14。

表 12-26 对应表

Correspondence Table				
品牌	收入水平			Active Margin
	低	中	高	
A	2	7	16	25
B	49	7	3	59
C	4	5	23	32
D	4	49	5	58
E	15	2	5	22
F	1	7	14	22
Active Margin	75	77	66	218

表 12-26 列出了所有参与分析的变量类别及其对应的观测值和边际和。用该表中数字可同原始资料做对比, 以确认数据录入的正确性。

表 12-27 汇总表

Summary								
Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation
1	.729	.531			.608	.608	.047	.165
2	.586	.343			.392	1.000	.058	
Total		.874	190.534	.000 ^a	1.000	1.000		

a. 10 degrees of freedom

表 12-28 品牌类别的坐标值

Overview Row Points ^a									
品牌	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		Total
					1	2	1	2	
A	.115	.620	-.731	.068	.061	.105	.472	.528	1.000
B	.271	-1.194	.153	.285	.530	.011	.987	.013	1.000
C	.147	.496	-1.047	.120	.049	.275	.218	.782	1.000
D	.266	.725	1.074	.282	.192	.524	.362	.638	1.000
E	.101	-.843	-.215	.055	.098	.008	.950	.050	1.000
F	.101	.708	-.672	.064	.069	.078	.580	.420	1.000
Active Total	1.000			.874	1.000	1.000			

a. Symmetrical normalization

表 12-27 同表 12-20 中显示的格式一样。从该表可见, 第一个维度的惯量值为 0.531、奇异值为 0.729, 第二个维度的惯量值为 0.343、奇异值为 0.586, 它们分别解释了总信息

量的 60.8%和 39.2%的信息，前两个维度累计解释了总信息量的 100%。因此，二维图形完全可以表示两个变量间的信息。列联表行列独立性的卡方检验结果为 $\chi^2=190.534$, $P=0.000<0.05$ ，它表明列联表的行列之间有较强的相关性。

表 12-28 的格式同表 12-21。左侧第一列为变量品牌的 6 个取值类别，第二列为其质量（Mass）值，也就是品牌的每个类别所占的百分比。第三列为每个品牌在两个维度中的分值，也就是在平面直角坐标系中的坐标值。第四列为每个品牌的惯量值。第五列为每个品牌对各个维度的贡献量，包括样品点对维度惯量的贡献和维度对样品点惯量的贡献两种。

品牌 A 的坐标为（0.620，-0.731），B 的坐标为（-1.194，0.153），C 的坐标为（0.496，-1.047），D 的坐标为（0.725，1.074），E 的坐标为（-0.843，-0.215，），F 的坐标为（0.708，-0.672）。

表 12-29 收入类别的坐标值

Overview Column Points ^a									
收入水平	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		Total
					1	2	1	2	
低	.344	-1.177	.051	.348	.654	.002	.999	.001	1.000
中	.353	.664	.847	.262	.214	.433	.433	.567	1.000
高	.303	.563	-1.046	.264	.132	.566	.265	.735	1.000
Active Total	1.000			.874	1.000	1.000			

a. Symmetrical normalization

表 12-29 的格式同表 12-28。区别是，它是描述收入类别的。
3 个收入点的坐标为：低（-1.177，0.051）、中（0.664，0.847）、高（0.563，-1.046）。

图 12-14 显示了由表 12-30 和表 12-31 中给出的各品牌类别和收入类别的坐标点所组成的对应分析点图。从表中可见，低收入人群倾向于选择品牌 B 和 E，中等收入人群倾向于选择品牌 D，而高收入人群倾向于选择品牌 A、F 和 C。

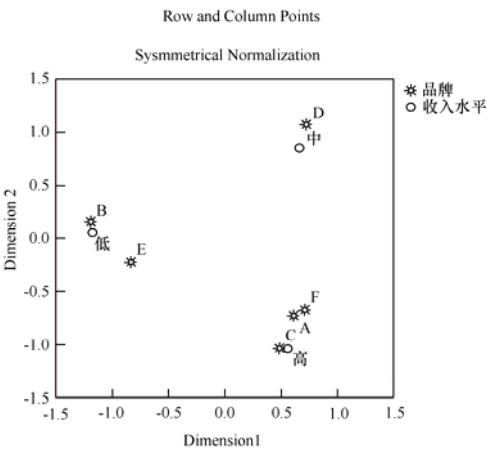


图 12-14 对应分析图

参 考 文 献

- [1] 卢纹岱等. SPSS for Windows 统计分析 (第 2 版). 北京: 电子工业出版社, 2002
- [2] 卢纹岱等. SPSS for Windows 统计分析 (第 3 版). 北京: 电子工业出版社, 2006
- [3] 宋志刚等. SPSS 16.0 实用教程. 北京: 人民邮电出版社, 2008
- [4] 张文彤. SPSS 11 统计分析教程. 北京: 北京希望电子出版社, 2002
- [5] Alan Agresti, Barbara Finlay. Statistical Methods for the Social Sciences (Fourth Edition). Prentice Hall, Inc. 2009
- [6] 王静龙, 梁小筠. 非参数统计分析. 北京: 高等教育出版社, 2006
- [7] 王静龙, 梁小筠. 定性数据统计分析. 北京: 中国统计出版社, 2008
- [8] 吴喜之. 非参数统计. 北京: 中国统计出版社, 1999
- [9] 吴喜之. 统计学: 从数据到结论. 北京: 中国统计出版社, 2004
- [10] 于秀林, 任雪松. 多元统计分析. 北京: 中国统计出版社, 1998
- [11] Nancy L. Leech, Karen C. Barrett, George A. Morgan 著, 何丽娟, 朱红兵译. SPSS 统计应用与解析 (第 3 版). 北京: 电子工业出版社, 2009
- [12] 清华大学应用数学系概率统计教研组编. 概率论与数量统计. 吉林: 吉林教育出版社, 1987
- [13] 茆诗松等. 回归分析及其试验设计. 上海: 华东师范大学出版社, 1981
- [14] 陈希孺, 倪国熙. 数理统计学教程. 上海: 上海科学技术出版社, 1988
- [15] 汪同三, 张涛等. 组合预测—理论、方法及应用. 北京: 社会科学文献出版社, 2008
- [16] W.G. 吉尔克里斯特. 统计预测. 北京: 机械工业出版社, 1984
- [17] 梁之舜等. 概率论及数理统计. 北京: 高等教育出版社, 1988
- [18] George E.P. Box, Gwilym M. Jenkins, Gregory C. Reinsel 著, 顾岚主译. 时间序列分析-预测与控制. 北京: 中国统计出版社, 1997
- [19] C. 查特菲尔德著, 骆振华译. 时间序列分析导论. 福建: 厦门大学出版社, 1987
- [20] H.L. 阿尔德, E.B. 罗斯勒著, 胡崇能等校译. 概率与统计导论. 北京: 北京大学出版社, 1984
- [21] 江天骥. 归纳逻辑导论. 湖南: 湖南人民出版社, 1987
- [22] 孙山泽. 抽样调查. 北京: 北京大学出版社, 2004
- [23] 陈家鼎等. 概率统计讲义. 北京: 高等教育出版社, 1980
- [24] 王保福等. 概率论及数理统计. 上海: 同济大学出版社, 1984
- [25] 金益. 试验设计与统计分析. 北京: 中国农业出版社, 2007
- [26] 吴诚欧等. 近代实用多元统计分析. 北京: 气象出版社, 2007
- [27] 杨振海, 张忠占. 应用数量统计. 北京: 北京工业大学出版社, 2005
- [28] 王岩等. 数理统计与 MATLAB 工程数据分析. 北京: 清华大学出版社, 2006
- [29] Stephen W. Raudenbush, Anthony S. Bryk 著, 郭志刚等译. 分层线性模型: 应用与数据分析方法. 北京: 社会科学文献出版社, 2007
- [30] 胡良平. 统计学三型理论在实验设计中的应用. 北京: 人民军医出版社, 2006
- [31] 韦博成. 近代非线性回归分析. 南京: 东南大学出版社, 1989
- [32] 付强. 数据处理方法及其农业应用. 北京: 科学出版社, 2006
- [33] 徐国祥. 统计预测和决策. 上海: 上海财经大学出版社, 1998
- [34] Frank R. Giordano, Maurice D. Weir, William P. Fox 著, 叶其孝, 姜启源等译. 数学建模 (原书第 3 版). 北京: 机械工业出版社, 2007
- [35] 朱秀娟, 洪再吉. 概率统计问答 150 题 (修订本). 湖南: 湖南科学技术出版社, 1985
- [36] 熊全淹, 叶明训. 线性代数 (第三版). 北京: 高等教育出版社, 1987
- [37] 李志辉等. SPSS for Windows 统计分析教程 (第 2 版). 北京: 电子工业出版社, 2006
- [38] 孙尚拱. 实用多变量统计方法与计算程序. 北京: 北京医科大学、中国协和医科大学联合出版社, 1990
- [39] 刘达民, 程岩. 应用统计. 北京: 化学工业出版社, 2004
- [40] 刘光祖. 概率论与应用数量统计. 北京: 高等教育出版社, 2000
- [41] 吴辉. 英汉统计词汇. 北京: 中国统计出版社, 1998
- [42] 郑家亨. 统计大辞典. 北京: 中国统计出版社, 1995